ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

# ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

# ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ ΜΕΙΩΣΗΣ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

***(APPLICATION OF DATA REDUCTION ALGORITHMS ON TIME SERIES CLASSIFICATION)***

**Του φοιτητή**

**Τατόγλου Χρήστου**

**Αρ. Μητρώου: 99/1391**

**Επιβλέπων καθηγητής**

**Καραμητόπουλος Λεωνίδας**

**Θεσσαλονίκη 2013**

Πτυχιακή εργασία του

**Χρήστου Τατόγλου**

xtatty@gmail.com

## ΠΡΟΛΟΓΟΣ

Η παρούσα πτυχιακή εργασία εκπονήθηκε στα πλαίσια των σπουδών μου στο τμήμα Πληροφορικής της Σχολής Τεχνολογικών Εφαρμογών (Σ.Τ.ΕΦ.) του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Θεσσαλονίκης (Α.Τ.Ε.Ι.Θ). Κατά την διάρκεια της εκπόνησής της μου δόθηκε η ευκαιρία να γνωρίσω ένα νέο αντικείμενο και να διευρύνω τις γνώσεις μου πάνω στον πολυδιάστατο κόσμο της πληροφορικής.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή αυτής της πτυχιακής εργασίας, Επιστημονικό Συνεργάτη του τμήματος Πληροφορικής του Α.Τ.Ε.Ι.Θ., κ. Λεωνίδα Καραμητόπουλο για την ουσιαστική καθοδήγηση, κατανόηση και την στήριξη του προκειμένου να ολοκληρωθεί η εργασία αυτή.

Επίσης ευχαριστώ πολύ τον φίλο μου και υποψήφιο Διδάκτορα του τμήματος Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας Στέφανο Ουγιάρογλου, για την πολύτιμη βοήθεια που μου προσέφερε και για την παραχώρηση της προγραμματιστικής υλοποίησης των αλγορίθμων που χρησιμοποιήθηκαν.

Κλείνοντας θέλω να αφιερώσω την πτυχιακή αυτή στη μνήμη των γονιών μου, Σπύρο και Ελένη, οι οποίοι με στήριξαν ηθικά και οικονομικά όσο ήταν εν ζωή. Η ξαφνική απουσία τους δυσκόλεψε το έργο μου, αλλά η γνωστή επιθυμία τους να ολοκληρώσω τον κύκλο σπουδών μου ήταν αυτή που μου έδωσε τη δύναμη να συνεχίσω και να φέρω σε πέρας αυτό το εγχείρημα.

## ΠΕΡΙΛΗΨΗ

Ένας αλγόριθμος κατηγοριοποίησης ή αλλιώς ένας κατηγοριοποιητής είναι μια τεχνική εξόρυξης δεδομένων, που προσπαθεί να κατηγοριοποιήσει δεδομένα σε ένα σύνολο προκαθορισμένων κλάσεων. Ένας κατηγοριοποιητής μπορεί να αξιολογηθεί από την ακρίβεια καθώς και την ταχύτητα στην κατηγοριοποίηση που επιτυγχάνει. Από το κριτήριο της ταχύτητας εξαρτάται το κατά πόσο ο κατηγοριοποιητής μπορεί να εφαρμοστεί σε μεγάλο όγκο δεδομένων.

Μια από τις πιο γνωστές μεθόδους κατηγοριοποίησης είναι αυτή των k εγγύτερων γειτόνων (k-Nearest Neighbors (k-NN)). Σε γενικές γραμμές, θεωρείται ότι είναι ένας απλός και αποτελεσματικός κατηγοριοποιητής που έχει πολλές εφαρμογές. Ωστόσο, η αναζήτηση πλησιέστερων γειτόνων είναι μια χρονοβόρα διαδικασία. Συνεπώς, η εφαρμογή του εν λόγω κατηγοριοποιητή δεν ενδείκνυται για μεγάλα σύνολα δεδομένων λόγω του υψηλού υπολογιστικού κόστους. Στη διεθνή βιβλιογραφία, προτείνονται τεχνικές μείωσης του όγκου δεδομένων για την επιτάχυνση της διαδικασίας αναζήτησης του k-NN κατηγοριοποιητή.

Οι τεχνικές μείωσης του όγκου των δεδομένων στοχεύουν στο να μειώσουν το υπολογιστικό κόστος της αναζήτησης των εγγύτερων γειτόνων όσο το δυνατόν περισσότερο και παράλληλα να διατηρήσουν την ακρίβεια κατηγοριοποίησης σε υψηλό επίπεδο. Στην παρούσα εργασία σκοπός είναι η συνοπτική παρουσίαση τέτοιου είδους τεχνικών καθώς και η διεξαγωγή πειραματικής μελέτης σε σύνολα δεδομένων χρονοσειρών. Η πειραματική μελέτη αποσκοπεί στην σύγκριση της αποτελεσματικότητας των τεχνικών σε τέτοιου είδους δεδομένα.

## ABSTRACT

A classification algorithm or a classifier is a data mining technique that attempts to map data to a set of predefined classes. A classifier can be evaluated by the accuracy achieved and speed of the classification. Speed is the criteria that defines the classifier's ability to work on large databases.

One of the most popular classification methods is the k-Nearest Neighbors (k-NN) classifier. In general, it is considered to be a simple and effective classifier which has many applications. However, the nearest neighbor search is a time consuming process. Therefore, the application of these classifiers is not suitable for large data sets because of the high computational cost. In the international literature, techniques are proposed in order to reduce the volume of data to speed up the search process of the k-NN classifier.

These methods aim to reduce the computational cost of the nearest neighbors as much as possible and at the same time to maintain the classification accuracy at a high level. This paper aims to summarize some of these techniques and to conduct an experimental study on data sets of time series. The objective is to compare the effectiveness of those techniques on such data sets.

# ΠΕΡΙΕΧΟΜΕΝΑ

## Ευρετήριο σχημάτων

## Ευρετήριο πινάκων

# ΕΙΣΑΓΩΓΗ

Εξόρυξη δεδομένων (data mining) είναι η διερεύνηση και η ανάλυση μεγάλων ποσοτήτων δεδομένων, με σκοπό την ανακάλυψη της γνώσης που πιθανότατα κρύβουν. Διάφοροι αλγόριθμοι εξόρυξης δεδομένων έχουν προταθεί στη βιβλιογραφία. Κάθε ένας υπάγεται σε μια κατηγορία αλγορίθμων ανάλογα με τη γνώση που καλείται να εξορύξει. Η παρούσα πτυχιακή εργασία επικεντρώνεται στην κατηγορία αλγορίθμων κατηγοριοποίησης ή απλά στους κατηγοριοποιητές (classifiers).

Οι κατηγοριοποιητές έχουν ως στόχο την πρόβλεψη της κατηγορίας (class) στην οποία ανήκουν μη-κατηγοριοποιημένα αντικείμενα (data items). Αυτό επιτυγχάνεται χρησιμοποιώντας αντικείμενα των οποίων η κατηγορία είναι γνωστή. Τα δεδομένα αυτά καλούνται δεδομένα εκπαίδευσης (training data) Ασφαλώς, η πρόβλεψη της κατηγορίας πρέπει να πραγματοποιείται με την υψηλότερη δυνατή ακρίβεια (accuracy).

Ο κατηγοριοποιητής k εγγύτερων γειτόνων (k-Nearest Neighbors classifier) είναι ένας από τους πιο γνωστούς κατηγοριοποιητές που συναντάμε στη βιβλιογραφία. Λειτουργεί ως εξής: αν κάποιο νέο αντικείμενο πρέπει να κατηγοριοποιηθεί, ο αλγόριθμος αναζητά και ανακτά τα k εγγύτερα αντικείμενα στο σύνολο δεδομένων εκπαίδευσης χρησιμοποιώντας ένα μέτρο απόστασης (distance metric). Στη συνέχεια, ο αλγόριθμος κατηγοριοποιεί το αντικείμενο στην πλειοψηφούσα κατηγορία, δηλαδή στην κατηγορία στην οποία ανήκουν οι περισσότεροι από τους k εγγύτερους γείτονες. Ασφαλώς, η απόδοση του κατηγοριοποιητή εξαρτάται από το μέτρο απόστασης που υιοθετείται καθώς και από την τιμή της παραμέτρου k. Μερικά χαρακτηριστικά που καθιστούν τον κατηγοριοποιητή k εγγύτερων γειτόνων δημοφιλή είναι τα εξής: (i) μπορεί να εφαρμοστεί σε πολλά πεδία, (ii) μπορεί να υλοποιηθεί πολύ εύκολα, (iii) η λειτουργία του μπορεί να κατανοηθεί εύκολα από το χρήστη και (iv), επιτυγχάνει υψηλή ακρίβεια.

Αν και η ακρίβεια που επιτυγχάνουν οι διάφοροι κατηγοριοποιητές είναι το σημαντικότερο κριτήριο αποτίμησης της απόδοσης τους, αυτό δεν είναι το μόνο. Η κατηγοριοποίηση ενός αντικειμένου πρέπει να εκτελείται γρήγορα και χωρίς την ανάγκη «κατανάλωσης» υψηλού υπολογιστικού κόστους. Έτσι, ένα δεύτερο και επίσης σημαντικό κριτήριο απόδοσης είναι το υπολογιστικό κόστος που απαιτούν.

Το κόστος που απαιτεί ο κατηγοριοποιητής των k εγγύτερων γειτόνων εξαρτάται από το πλήθος των δεδομένων εκπαίδευσης. Για την κατηγοριοποίηση ενός νέου αντικειμένου, υπολογίζει τόσες αποστάσεις όσο και το πλήθος των δεδομένων εκπαίδευσης. Γίνεται αντιληπτό ότι το υπολογιστικό κόστος του είναι συνήθως πολύ υψηλό, ενώ σε περιπτώσεις μεγάλων συνόλων δεδομένων, το τεράστιο κόστος επεξεργασίας που απαιτεί καθιστά την εκτέλεση του απαγορευτική. Αυτό αποτελεί ένα μειονέκτημα του αλγόριθμου.

Επίσης, ο εν λόγω κατηγοριοποιητής δεν προεπεξεργάζεται τα διαθέσιμα δεδομένα εκπαίδευσης για να κατασκευάσει ένα μοντέλο κατηγοριοποίησης, το οποίο στη συνέχεια θα χρησιμοποιηθεί για την εκτέλεση της κατηγοριοποίησης. Οπότε, τα δεδομένα εκπαίδευσης πρέπει πάντα να είναι διαθέσιμά προς προσπέλαση. Συνεπώς, ένα δεύτερο μειονέκτημα είναι η υψηλές απαιτήσεις αποθηκευτικού χώρου για την αποθήκευση των δεδομένων εκπαίδευσης.

Τα μειονεκτήματα του κατηγοριοποιητή των k εγγύτερων γειτόνων αποτελούν έναν ενεργό χώρο έρευνας εδώ και πολλά χρόνια και έχει προσελκύσει το ενδιαφέρον πολλών ερευνητών διαφόρων πεδίων της επιστήμης της πληροφορικής. Αποτέλεσμα είναι η παραγωγή διάφορων αλγορίθμων και τεχνικών που έχουν ως στόχο την επιτάχυνση της διαδικασίας αναζήτησης εγγύτερων γειτόνων ή/και την μείωση των απαιτήσεων σε χώρο αποθήκευσης. Παράλληλος στόχος είναι η διατήρηση της ακρίβειας σε υψηλά επίπεδα.

Η παρούσα πτυχιακή εργασία επικεντρώνεται στις τεχνικές που έχουν προταθεί για τη μείωση του όγκου των δεδομένων εκπαίδευσης με στόχο τη μείωση του πλήθους των αποστάσεων που πρέπει να υπολογιστούν (και συνεπώς την πιο γρήγορη κατηγοριοποίηση εγγύτερων γειτόνων) και την μείωση των απαιτήσεων για αποθηκευτικό χώρο. Επίσης, η εργασία επικεντρώνεται σε δεδομένα χρονοσειρών (time series), δηλαδή αντικείμενα που συνθέτονται από την παρακολούθηση μιας κατάστασης σε διαφορετικές χρονικές στιγμές. Οι διάφοροι αλγόριθμοι μείωσης (ή συμπύκνωσης) των δεδομένων εκπαίδευσης (Data Reduction Techniques) που συναντάμε στη βιβλιογραφία δεν έχουν εφαρμοστεί σε τέτοιου είδους δεδομένα. Αυτό αποτελεί το κίνητρο για την εκπόνηση της εργασίας. Η συνεισφορά της παρούσας εργασίας είναι η διεξαγωγή μιας εκτεταμένης πειραματικής συγκριτικής μελέτης και η εξαγωγή των αντίστοιχων συμπερασμάτων αναφορικά με την απόδοση γνωστών αλγορίθμων μείωσης των δεδομένων εκπαίδευσης σε σύνολα χρονοσειρών.

Το περιεχόμενο της παρούσας εργασίας έχει οργανωθεί σε τέσσερα κεφάλαια. Στο πρώτο κεφάλαιο, αρχικά, γίνεται μια μικρή εισαγωγή σε έννοιες που σχετίζονται με την εξόρυξη δεδομένων. Στη συνέχεια, το κεφάλαιο αυτό, επικεντρώνεται στην παρουσίαση εννοιών που σχετίζονται με τους αλγορίθμους κατηγοριοποίησης. Το περιεχόμενο του δεύτερου κεφαλαίου αφιερώνεται αποκλειστικά στην αναλυτική παρουσίαση της λειτουργίας και των χαρακτηριστικών του κατηγοριοποιητή k εγγύτερων γειτόνων. Επίσης, το κεφάλαιο παρουσιάζει τα πλεονεκτήματα και τα μειονεκτήματα του κατηγοριοποιητή, ενώ γίνεται αναφορά στις μεθόδους που έχουν προταθεί για την εξάλειψη των μειονεκτημάτων. Το τρίτο κεφάλαιο αρχικά παρουσιάζει έννοιες που σχετίζονται με την μείωση του όγκου των δεδομένων. Στη συνέχεια, το κεφάλαιο παρουσιάζει τη λειτουργία και τα χαρακτηριστικά γνωστών τεχνικών μείωσης δεδομένων εκπαίδευσης. Στο τέλος του κεφαλαίου παρουσιάζονται έννοιες που σχετίζονται με τη μείωση διαστάσεων δεδομένων χρονοσειρών. Όλες οι τεχνικές που παρουσιάζονται στο τρίτο κεφάλαιο, χρησιμοποιήθηκαν κατά τη διάρκεια εκπόνησης της πειραματικής συγκριτικής μελέτης. Τέλος, το τελευταίο κεφάλαιο, αφού πρώτα αναλύει το περιβάλλον όπου εκτελέστηκαν όλα τα πειράματα,

παρουσιάζει τα αποτελέσματα της πειραματικής μελέτης με τα αντίστοιχα συμπεράσματα.

# ΚΕΦΑΛΑΙΟ 1

# ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

## ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή στο ζήτημα της κατηγοριοποίησης (classification), το οποίο εμφανίζετε σε πολλά ερευνητικά πεδία της πληροφορικής. Αρχικά, παρουσιάζονται κάποια εισαγωγικά θέματα γύρω από την έννοια της κατηγοριοποίησης, ενώ στην συνέχεια παρουσιάζονται ορισμένοι από τους γνωστούς αλγόριθμους που χρησιμοποιούνται για την επίλυση τέτοιου είδους προβλημάτων καθώς και κάποια παραδείγματα.

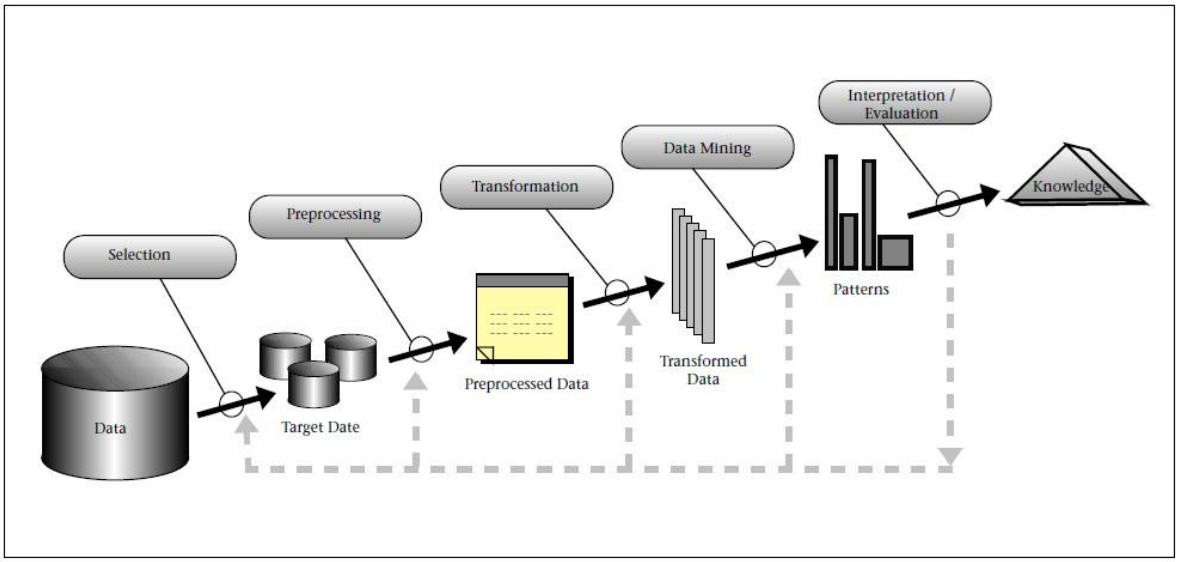
## 1.1 ΕΞΟΡΥΞΗ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΔΕΔΟΜΕΝΑ (DATA MINING)

Οι τεχνικές εξόρυξης δεδομένων (Han, 2011, Dunham, 2003) έχουν σκοπό την ανακάλυψη ενδιαφερόντων ή πρότυπων σχημάτων μέσα σε μεγάλα σύνολα δεδομένων, ώστε αυτά να βοηθήσουν τους ειδικούς να λαμβάνουν αποφάσεις σχετικά με σημαντικές, μελλοντικές δραστηριότητες. Η εξόρυξη γνώσης είναι στενά συνδεδεμένη με τις περιοχές των Βάσεων Δεδομένων, της Στατιστικής και της Τεχνητής Νοημοσύνης (Μηχανική μάθηση).

Κάποιος μπορεί να πει ότι όταν εκτελούμε ένα ερώτημα διατυπωμένο σε SQL, στην ουσία κάνουμε εξόρυξη δεδομένων. Τα πράγματα δεν είναι έτσι. Στην πραγματικότητα, όταν εκτελούμε ένα ερώτημα SQL γνωρίζουμε το τι αναζητούμε. Αντίθετα, εκτελώντας διαδικασίες εξόρυξης δεδομένων προσπαθούμε να ανακαλύψουμε την πιθανή γνώση που κρύβουν τα δεδομένα. Ασφαλώς, δεν γνωρίσουμε εκ των προτέρων το είδος της γνώσης που θα ανακαλυφθεί.

Όλες οι τεχνικές εξόρυξης δεδομένων εφαρμόζονται για να ικανοποιήσουμε σύνθετα αιτήματα, τα οποία καθορίζονται σε υψηλό επίπεδο με μερικές παραμέτρους που προσδιορίζονται από τον χρήστη. Η ικανοποίηση αυτών των αιτημάτων γίνεται χρησιμοποιώντας εξειδικευμένους αλγορίθμους. Ωστόσο, ένα τέτοιο αίτημα δεν ικανοποιείται απλά εφαρμόζοντας ένα τέτοιο αλγόριθμο. Συγκεκριμένα, οι διαδικασίες εξόρυξης γνώσης σε Βάσεις Δεδομένων (ή διαδικασίες KDD) απαιτούν τέσσερα στάδια. Το πρώτο στάδιο είναι η επιλογή των δεδομένων, όπου επιλέγεται το σύνολο δεδομένων και τα χαρακτηριστικά (attributes) που μας ενδιαφέρουν σε σχέση με το στόχο μας. Το δεύτερο στάδιο καλείται προεπεξεργασία των δεδομένων. Στο στάδιο αυτό, απομακρύνουμε το θόρυβο, χειριζόμαστε τις κενές τιμές, μετασχηματίζουμε τις τιμές των χαρακτηριστικών σε κοινές μονάδες μέτρησης και δημιουργούμε νέα χαρακτηριστικά συνδυάζοντας τα ήδη υπάρχοντα. Συχνά στη βιβλιογραφία, το στάδιο της προεπεξεργασίας χωρίζεται σε δυο επιμέρους σταδία: (i) της

προεπεξεργασίας και (ii) του μετασχηματισμού των δεδομένων. Το επόμενο στάδιο είναι το στάδιο της εξόρυξης γνώσης οπού εφαρμόζεται ο αλγόριθμος εξόρυξης γνώσης και εξάγουμε τα πραγματικά πρότυπα σχήματα. Το τέταρτο και τελευταίο στάδιο είναι αυτό της αξιολόγησης, όπου τα αποτελέσματα που προέκυψαν από την όλη διαδικασία ερμηνεύονται και αξιολογούνται από το χρήστη και προκύπτει η τελική γνώση. Το σχήμα 1, παρουσιάζει συνοπτικά την διαδικασία εξόρυξης γνώσης από βάσεις δεδομένων.




Σχήμα 1 "Η διαδικασία KDD (Knowledge Discovery in Databases)"

Σημαντικές κατηγορίες αλγόριθμων Εξόρυξης δεδομένων είναι οι εξής:

- **Συσταδοποίηση (Clustering):** η οργάνωση μιας συλλογής από αντικείμενα (instances) σε συστάδες (clusters) με βάση κάποιο μέτρο ομοιότητας – Αναζήτηση και Ανάλυση έκτοπων (Outliers)
- **Κατηγοριοποίηση (Classification):** Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κατηγοριών ή κλάσεων
- **Κανόνες Συσχέτισης (Association rules):** Ανακάλυψη κρυμμένων «συσχετίσεων» που υπάρχουν μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων
- **Πρότυπα Ακολουθιών (Sequential Patterns):** Ανακάλυψη των πιο συχνά εμφανιζόμενων προτύπων σχετικά με το χρόνο (χρονοσειρές) ή άλλες ακολουθίες (κείμενα, μουσικές νότες, δεδομένα καιρού, δεδομένα χρηματιστηρίου, ακολουθίες DNA)

Η εργασία αυτή επικεντρώνεται σε έννοιες που σχετίζονται με τους αλγορίθμους κατηγοριοποίησης και συγκεκριμένα σε αυτούς που βασίζονται σε μέτρα απόστασης.

## 1.2 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Η **κατηγοριοποίηση (classification)** είναι η πιο γνωστή και πιο δημοφιλής τεχνική **εξόρυξης γνώσης**. Πολλές εταιρίες του ιδιωτικού και του δημόσιου τομέα χρησιμοποιούν σε καθημερινή βάση συστήματα κατηγοριοποίησης. Παραδείγματα τέτοιου είδους συστημάτων είναι τα συστήματα αναγνώρισης προτύπων, συστήματα ιατρικών διαγνώσεων, συστήματα έγκρισης δανείων και πιστωτικών καρτών, συστήματα ανίχνευσης λαθών σε βιομηχανικές εφαρμογές, συστήματα κατηγοριοποίησης των τάσεων στην οικονομία κ.α. Για παράδειγμα όταν κάποιος προβλέπει μια ηλικία, στην ουσία επιλύει ένα πρόβλημα κατηγοριοποίησης.

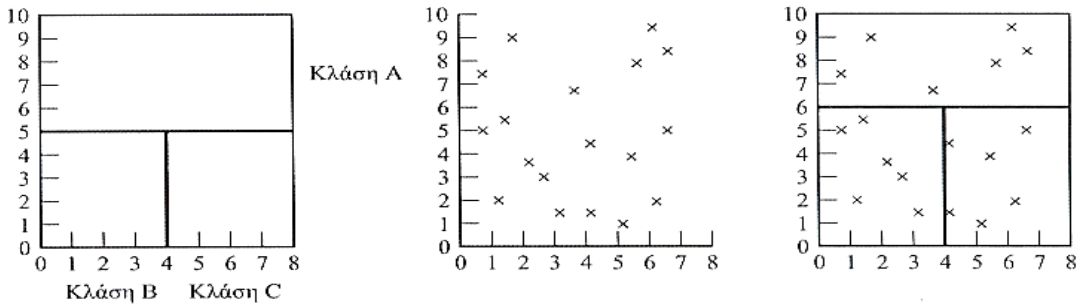
Όλες οι προσεγγίσεις στην εκτέλεση της κατηγοριοποίησης προϋποθέτουν γνώση των δεδομένων. Συνήθως χρησιμοποιούμε ένα σύνολο εκπαίδευσης για να καθορίσει τις συγκεκριμένες παραμέτρους που απαιτούνται από την τεχνική. Τα δεδομένα εκπαίδευσης (training data) αποτελούνται από ένα δείγμα δεδομένων εισόδου καθώς επίσης και από την κατηγοριοποίηση που έχει δοθεί σε αυτά τα δεδομένα. Μπορούμε να ορίσουμε το πρόβλημα της κατηγοριοποίησης ως ακολούθως: Η κατηγοριοποίηση (classification), είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες. Τις ομάδες αυτές συχνά τις καλούμε κατηγορίες ή κλάσεις. Ο ορισμός αυτός θεωρεί την κατηγοριοποίηση σαν μια απεικόνιση από τη Βάση Δεδομένων στο σύνολο των κατηγοριών. Οι κατηγορίες είναι προκαθορισμένες, δεν επικαλύπτονται και διαμερίζουν ολόκληρη την Βάση Δεδομένων. Κάθε στοιχείο της Βάσης Δεδομένων τοποθετείται σε ακριβώς μια κατηγορία.

Η επίλυση των προβλημάτων κατηγοριοποίησης περιλαμβάνει δύο βασικά στάδια (σχήμα 3): Αρχικά, δημιουργούμε ένα μοντέλο από την αξιολόγηση και την ανάλυση των δεδομένων εκπαίδευσης. Αυτό το βήμα έχει σαν είσοδο τα δεδομένα εκπαίδευσης και σαν έξοδο ένα ορισμό του μοντέλου που αναπτύχθηκε. Το μοντέλο που δημιουργείται από αυτό το στάδιο είναι σε θέση να κατηγοριοποιεί τα δεδομένα εκπαίδευσης με όσο το δυνατό μεγαλύτερη ακρίβεια. Όταν είναι ήδη γνωστές οι κατηγορίες του συνόλου των δεδομένων εκπαίδευσης, δηλαδή το σύνολο των δεδομένων εκπαίδευσης περιλαμβάνει ένα χαρακτηριστικό το οποίο δείχνει την κλάση στην οποία κατηγοριοποιείται το κάθε αντικείμενο (data item), τότε το βήμα αυτό καλείται εποπτευμένη μάθηση (supervised learning).Σε αντίθετη περίπτωση, δηλαδή όταν δεν είναι γνωστές οι κατηγορίες του συνόλου των δεδομένων εκπαίδευσης, τότε το βήμα αυτό καλείται μη εποπτευμένη μάθηση (unsupervised learning - clustering). Στην εργασία αυτή δεν εξετάζεται η μη εποπτευόμενη μάθηση. Αντίθετα επικεντρωνόμαστε στην εποπτευμένη μάθηση, δηλαδή στην κατηγοριοποίηση. Στην συνέχεια εφαρμόζουμε το μοντέλο που αναπτύχθηκε στο προηγούμενο βήμα κατηγοριοποιώντας τα αντικείμενα της υπό εξέταση Βάσης Δεδομένων (μελλοντικές περιπτώσεις)

Εάν και το δεύτερο βήμα στην πραγματικότητα εκτελεί την κατηγοριοποίηση, η περισσότερη έρευνα στον χώρο έχει γίνει για το πρώτο βήμα. Το δεύτερο βήμα είναι συνήθως εύκολο στην υλοποίηση.

Αξίζει να σημειωθεί ότι αντίθετα με τις μεθόδους που αναπτύσσουν ένα μοντέλο κατηγοριοποίησης (eager classifiers), υπάρχουν μέθοδοι κατηγοριοποίησης οι οποίοι δε δημιουργούν κάποιο μοντέλο. Στις περιπτώσεις αυτές μοντέλο κατηγοριοποίησης αποτελεί το σύνολο δεδομένων εκπαίδευσης. Οι μέθοδοι αυτοί καλούνται lazy κατηγοριοποιητές και βασίζονται στην εξέταση του συνόλου εκπαίδευσης την στιγμή της κατηγοριοποίησης. Χαρακτηριστικό παράδειγμα των lazy κατηγοριοποιητών είναι η μέθοδος των k εγγύτερων γειτόνων.

Το πρόβλημα της κατηγοριοποίησης γίνεται ευκολότερα κατανοητό μελετώντας το σχήμα 2. Ας υποθέσουμε ότι μας δίνεται μια Βάση Δεδομένων που αποτελείται από αντικείμενα της μορφής $t = <x,y>$ όπου $0 \le x \le 8$ και $0 \le y \le 10$. Το σχήμα 2 (α) παρουσιάζει τις προκαθορισμένες κατηγορίες – κλάσεις. Το σχήμα 2 (β) εμφανίζει τα δείγματα δεδομένων εισόδου. Τέλος, το σχήμα 2 (γ) παρουσιάζει την κατηγοριοποίηση των δεδομένων με βάση τις ορισμένες κατηγορίες.




Σχήμα 2 "Το πρόβλημα της κατηγοριοποίησης"

Ένα πολύ σημαντικό ζήτημα σχετικό με την κατηγοριοποίηση είναι η υπερπροσαρμογή. Συγκεκριμένα, λέγοντας υπερπροσαρμογή εννοούμε το φαινόμενο κατά το οποίο η τεχνική κατηγοριοποίησης ταιριάζει ακριβώς τα δεδομένα εκπαίδευσης και ίσως να μη μπορεί να εφαρμοστεί σε πιο ευρύ πληθυσμό δεδομένων. Για παράδειγμα, ας υποθέσουμε ότι τα δεδομένα εκπαίδευσης περιέχουν λανθασμένα δεδομένα ή δεδομένα με θόρυβο. Σε αυτή την περίπτωση, το ακριβές ταίριασμα των δεδομένων δεν είναι επιθυμητό.

## 1.3 ΑΠΟΔΟΣΗ ΤΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Η απόδοση των αλγορίθμων συνήθως εξετάζεται με την εκτίμηση της ακρίβειας (accuracy) της κατηγοριοποίησης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Η εκτίμηση της ακρίβειας είναι ένα πολύ σημαντικό ζήτημα στο χώρο της κατηγοριοποίησης, αφού κάτι τέτοιο μας δείχνει το πόσο καλά ανταποκρίνεται ο αλγόριθμος μας για δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Η εκτίμηση της ακρίβειας είναι επίσης θεμιτή αφού μας επιτρέπει την σύγκριση των διαφόρων αλγορίθμων κατηγοριοποίησης.

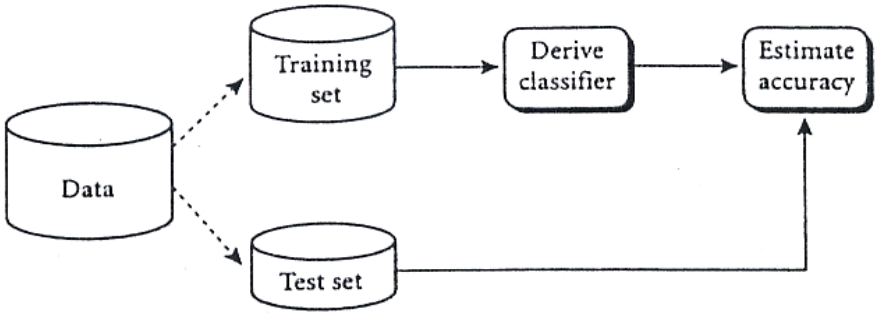
Αν και η ακρίβεια είναι το πιο σημαντικό μέτρο αποτίμησης της απόδοσης του αλγορίθμου κατηγοριοποίησης που χρησιμοποιούμε, υπάρχουν και άλλα μέτρα σύγκρισης:

- **Ταχύτητα:** Κόστος υπολογισμού (συμπεριλαμβανομένου την παραγωγή και τη χρήση του μοντέλου)
- **Rebustness:** Σωστή πρόβλεψη με ελλιπή δεδομένα ή δεδομένα με θόρυβο
- **Scalability:** Αποδοτική κατασκευή του μοντέλου δοθέντος μεγάλη ποσότητα δεδομένων
- **Interpretability:** Επίπεδο κατανόησης και γνώση που παρέχεται από το μοντέλο. (Μπορεί να εκτιμηθεί μετρώντας το πόσο πολύπλοκο είναι το μοντέλο π.χ. αριθμός κόμβων στα δένδρα απόφασης, αριθμός επιπέδων στα νευρωνικά δίκτυα κ.α.)

Τώρα ας επιστρέψουμε στο σημαντικότερο μέτρο μέτρησης απόδοσης, δηλαδή την ακρίβεια στη πρόβλεψη της κλάσης. Το μέτρο αυτό είναι το πιο σημαντικό, ωστόσο δε θα πρέπει να υπολογίζεται ανεξάρτητα από τα υπόλοιπα μέτρα. Για παράδειγμα, δεν έχει νόημα το να έχουμε έναν αλγόριθμο κατηγοριοποίησης που μας δίνει αποτελέσματα με πολύ υψηλή ακρίβεια μετά από πολύ χρόνο. Ίσως να ήταν καλύτερη επιλογή το να έχουμε έναν αλγόριθμο κατηγοριοποίησης που να μας δίνει αποτελέσματα με λίγο χαμηλότερη ακρίβεια από τον καλύτερο (ως προς την ακρίβεια) αλγόριθμο κατηγοριοποίησης, αλλά πιο σύντομα. Η ακρίβεια της κατηγοριοποίησης συνήθως υπολογίζεται με τον καθορισμό του ποσοστού των αντικειμένων που τοποθετούνται την σωστή κατηγορία.

Υπάρχουν τρεις τρόποι που μας επιτρέπουν να εκτιμήσουμε την ακρίβεια του αλγορίθμου κατηγοριοποίησης: Μπορούμε να χρησιμοποιήσουμε ένα σύνολο δεδομένων αρχικά για να εκπαιδεύσουμε τον αλγόριθμο μας και στην συνέχεια να χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων για να εκτιμήσουμε την ακρίβεια του αλγορίθμου. Μια τέτοια επιλογή θα μας οδηγούσε σε μια πολύ αισιόδοξη εκτίμηση της ακρίβειας, αφού ο αλγόριθμος εκπαιδεύεται αλλά και δοκιμάζεται με το ίδιο σύνολο δεδομένων.

Άλλος ένας τρόπος εκτίμησης της ακρίβειας ενός αλγορίθμου κατηγοριοποίησης είναι η μέθοδο της κατακράτησης (holdout method) – (Σχήμα 3). Χρησιμοποιώντας αυτή την μέθοδο, το σύνολο δεδομένων που έχουμε στην διάθεση μας, χωρίζεται με τυχαίο τρόπο σε δυο ανεξάρτητα σύνολα δεδομένων. Το πρώτο ονομάζεται σύνολο δεδομένων εκπαίδευσης (training set) και χρησιμοποιείται για την εκπαίδευση του αλγορίθμου κατηγοριοποίησης. Το δεύτερο ονομάζεται σύνολο δεδομένων δοκιμής (testing set) που χρησιμοποιείται για την δοκιμή του αλγορίθμου και την εκτίμηση της ακρίβειας. Στις περισσότερες περιπτώσεις, χρησιμοποιούνται τα 2/3 του συνόλου δεδομένων σαν σύνολο εκπαίδευσης και το υπόλοιπο 1/3 σαν σύνολο δοκιμής.




Σχήμα 3 "Εκτίμηση ακρίβειας χρησιμοποιώντας την μέθοδο της κατακράτησης"

Ένας λίγο πιο σύνθετος τρόπος εκτίμησης της απόδοσης είναι το **cross validation**. Σε αυτή την περίπτωση, για πιο ασφαλή αποτελέσματα, ο διαχωρισμός των δεδομένων σε σύνολο εκπαίδευσης και δοκιμής γίνεται πολλές φορές χρησιμοποιώντας διαφορετικά υποσύνολα κάθε φορά και στο τέλος βγαίνει ο μέσος όρος από όλες τις επαναλήψεις. Υπάρχουν 2 παραλλαγές της μεθόδου cross validation:

- **K-fold cross-validation:** Σε αυτή τη μέθοδο το αρχικό σύνολο δεδομένων χωρίζεται σε k υποσύνολα. Από τα k υποσύνολα ένα χρησιμοποιείται για επαλήθευση και τα υπόλοιπα (k-1) υποσύνολα χρησιμοποιούνται για την εκπαίδευση. Η διαδικασία αυτή επαναλαμβάνεται k φορές (όσες και τα folds), όπου κάθε ένα από τα k υποσύνολα χρησιμοποιείται μία φορά σαν δεδομένα επαλήθευσης. Το μεγαλύτερο πλεονέκτημα αυτής της μεθόδου είναι ότι όλα τα αντικείμενα του συνόλου δεδομένων χρησιμοποιούνται και για εκπαίδευση αλλά και για επαλήθευση. Όπως γίνεται εύκολα κατανοητό, αυτή η προσέγγιση απαιτεί $k$ φορές περισσότερο χρόνο από την μέθοδο της κατακράτησης. Οι περισσότερες ερευνητικές εργασίες χρησιμοποιούν το 5 ή 10 σαν καταλληλότερος αριθμό επαναλήψεων.
- **Leave-one-out cross-validation:** Όπως δηλώνει και το όνομα η μέθοδος leave-one-out cross-validation (LOOCV), χρησιμοποιεί ένα μόνο αντικείμενο από το αρχικό σύνολο δεδομένων για επαλήθευση και όλα τα υπόλοιπα αντικείμενα χρησιμοποιούνται για εκπαίδευση. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να χρησιμοποιηθούν όλα τα αντικείμενα από μία τουλάχιστον φορά για επαλήθευση. Η διαδικασία είναι ίδια με αυτή του K-fold cross-validation, απλά στην προκειμένη περίπτωση ο αριθμός K (folds) είναι ίσος με τον αριθμό των αντικειμένων. Η μέθοδος leave-one-out συνήθως δίνει τα καλύτερα αποτελέσματα, αλλά έχει μεγάλο κόστος σε υπολογιστική ισχύ λόγω των πολλών επαναλήψεων που απαιτούνται για την ολοκλήρωση της εκπαίδευσης.

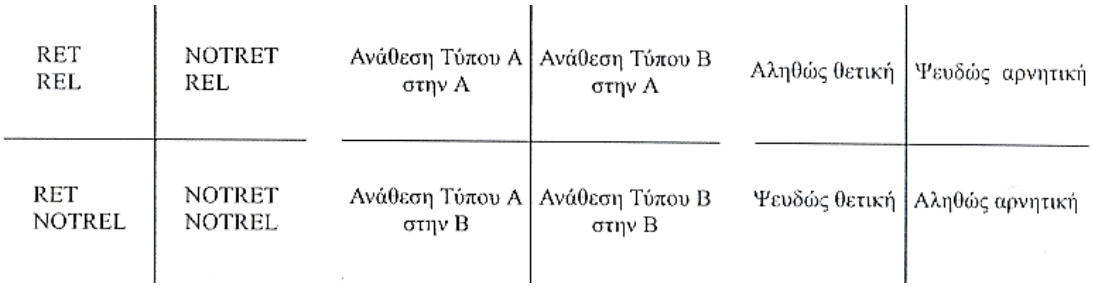
Στα πλαίσια της εκπόνησης της πειραματικής μελέτης που παρουσιάζεται σε αυτή την εργασία χρησιμοποιήθηκε η μέθοδος 5-fold cross validation

Επιπρόσθετα, ένας κατηγοριοποιητής μπορεί να αξιολογηθεί βάσει του κατά πόσο μπορεί να προβλέψει με υψηλή ακρίβεια τις σπάνιες κλάσεις. Για παράδειγμα είναι σημαντικότερο το να προβλεφθεί ένα σπάνιο και ακραίο καιρικό φαινόμενο και τελικά αυτό να μην επιβεβαιωθεί, παρά να μην προβλεφθεί και να πραγματοποιηθεί. Έστω ότι θέλουμε να εκπαιδεύσουμε έναν αλγόριθμο κατηγοριοποίησης με ένα σύνολο μετεωρολογικών δεδομένων, ώστε αυτός να είναι σε θέση να κατατάσσει τα μελλοντικά αντικείμενα στις κατηγορίες «ακραίο φαινόμενο» ή «μη ακραίο φαινόμενο». Μια εκτίμηση της ακριβείας γύρω στο 95% μπορεί να παρουσιάζει τον αλγόριθμο μας αρκετά ακριβή, ωστόσο τι γίνεται αν μόνο το 3-4% των δεδομένων εκπαίδευσης ανήκει στην κατηγορία «ακραίο φαινόμενο». Σε μια τέτοια περίπτωση, ένας αλγόριθμος κατηγοριοποίησης με ακρίβεια 95% ίσως να μην είναι δεκτός αφού στην πραγματικότητα θα μπορεί να αναγνωρίζει και να κατηγοριοποιεί μόνο τα αντικείμενα που ανήκουν στην κατηγορία «μη ακραίο φαινόμενο». Αντίθετα εμείς θέλουμε κατηγοριοποιητή που να είναι σε θέση να αναγνωρίζει τα «ακραίο φαινόμενο» αντικείμενα (positive samples) με υψηλή ακρίβεια. Για να γίνει αυτό μπορούμε να χρησιμοποιήσουμε τα μέτρα sensitivity και specificity αντίστοιχα:

$$sensitivity = \frac{t\_pos}{pos}, \quad specificity = \frac{t\_neg}{neg}$$

$t\_pos$ είναι ο αριθμός των true positives (αντικείμενα «ακραίων φαινομένων» που σωστά κατηγοριοποιήθηκαν σαν ακραία φαινόμενα), $pos$ είναι ο αριθμός των positive («ακραίων φαινομένων») αντικειμένων, $t\_neg$ είναι ο αριθμός των true negatives («μη ακραίων φαινομένων») αντικειμένων που σωστά κατηγοριοποιήθηκαν σαν «μη ακραίων φαινομένων», $neg$ είναι ο αριθμός των negative («ακραίων φαινομένων») αντικειμένων και $f\_pos$ είναι ο αριθμός των false positives («μη ακραίων φαινομένων» αντικειμένων που λανθασμένα κατηγοριοποιήθηκαν σαν «ακραία φαινόμενα»).

Επίσης, μπορούμε να εξετάσουμε την απόδοση της κατηγοριοποίησης με τρόπο όμοιο με αυτό που εφαρμόζεται στα συστήματα ανάκτησης πληροφοριών. Όταν έχουμε δύο κλάσεις, υπάρχουν τέσσερα πιθανά ενδεχόμενα κατηγοριοποίησης, όπως φαίνεται και στο σχήμα 4. Το πάνω αριστερά και κάτω δεξιά τεταρτημόριο υποδηλώνουν λανθασμένες ενέργειες. Η επίδοση της κατηγοριοποίησης θα μπορούσε να καθοριστεί με την απόδοση κάποιου κόστους σε κάθε ένα από τα τεταρτημόρια. Ωστόσο κάτι τέτοιο θα ήταν δύσκολο αφού θα χρειαζόντουσαν $m^2$ κόστη, όπου m είναι ο αριθμός των κατηγοριών.




α. Ανάκτηση πληροφοριών β. κατηγοριοποίηση στην κλάση Α γ. πρόβλεψη κατηγορίας

**Σχήμα 4 "Σύγκριση της απόδοσης της κατηγοριοποίησης με την ανάκτηση πληροφορίας"**

Με δεδομένη μια συγκεκριμένη κατηγορία $C_j$ και ένα αντικείμενο της Βάσης Δεδομένων $t_i$, αυτό το αντικείμενο είτε θα καταχωρηθεί σε αυτή την κατηγορία είτε όχι, ενώ στην πραγματικότητα μπορεί να είναι ή να μην είναι μέλος αυτής της κατηγορίας. Αυτή η παρατήρηση πάλι μας δίνει τα τέσσερα τεταρτημόρια που παρουσιάζονται στο σχήμα 4 (γ), τα οποία μπορούν να περιγραφούν με τους παρακάτω τρόπους:

- **Αληθώς θετικό (True Positive - TP):** το $t_i$ εκτιμάται ότι ανήκει στην κατηγορία $C_j$ και πράγματι ανήκει σε αυτήν
- **Ψευδώς θετικό (False Positive - FP):** το $t_i$ εκτιμάται ότι ανήκει στην κατηγορία $C_j$ ενώ στην πραγματικότητα δεν ανήκει σε αυτήν.
- **Αληθώς αρνητικό (True Negative - TN):** το $t_i$ εκτιμάται ότι δεν ανήκει στην κατηγορία $C_j$ και πράγματι δεν ανήκει σε αυτήν
- **Ψευδώς αρνητικό (False Negative - FN):** το $t_i$ εκτιμάται ότι δεν ανήκει στην κατηγορία $C_j$ ενώ στην πραγματικότητα ανήκει σε αυτήν.

Τέλος, ένας τρόπος παρουσίασης αποτελεσμάτων που επιδεικνύει την ακρίβεια της λύσης σε ένα πρόβλημα κατηγοριοποίησης είναι ο πίνακας ή μήτρα σύγχυσης (confusion matrix). Με δεδομένες m κατηγορίες μια μήτρα σύγχυσης είναι μια m x m μήτρα όπου κάθε καταχώρηση $C_{i,j}$ δείχνει τον αριθμό των αντικειμένων που τοποθετήθηκαν στην κατηγορία $C_j$ αλλά των οποίων η πραγματική κατηγορία είναι η $C_i$. Όπως καταλαβαίνουμε, οι καλύτερες λύσεις θα έχουν μόνο μηδενικές τιμές έξω από την κύρια διαγώνιο. Μια μήτρα σύγχυσης για τρεις κατηγορίες παρουσιάζεται στον πίνακα 1.

Πίνακας 1 "Μήτρα σύγχυσης"

| Πραγματική κατηγορία | Εκχώρηση | | |
|---|---|---|---|
| | **Short** | **Medium** | **Tall** |
| **Short** | 5 | 0 | 0 |
| **Medium** | 0 | 4 | 3 |
| **Tall** | 0 | 1 | 2 |

## 1.4 ΠΑΡΑΔΕΙΓΜΑΤΑ ΕΦΑΡΜΟΓΩΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Η ζωή μας είναι γεμάτη με παραδείγματα κατηγοριοποίησης. Στην παράγραφο αυτή, για να γίνει καλύτερα κατανοητή η έννοια της κατηγοριοποίησης, παρουσιάζονται κάποια παραδείγματα όπου θα μπορούσαν να εφαρμοστούν οι διάφοροι αλγόριθμοι.

**Παράδειγμα:** Οι εταιρείες πιστωτικών καρτών πρέπει να καθορίζουν, αν θα εγκρίνουν αγορές μέσω πιστωτικών καρτών. Ας υποθέσουμε ότι με βάση το αγοραστικό ιστορικό ενός πελάτη κάθε αγορά τοποθετείται σε μια από τις τέσσερις κατηγορίες: 1. Να εγκριθεί, 2. Να ζητηθούν επιπλέον στοιχεία ταυτότητας πριν την έγκριση, 3. Να μην εγκριθεί, 4. Να μην εγκριθεί και να ενημερωθεί και η αστυνομία.

Οι λειτουργίες εξόρυξης γνώσης εξυπηρετούν δύο σκοπούς. Τα δεδομένα του ιστορικού των πελατών πρέπει να εξεταστούν, ώστε να καθοριστεί πώς αυτά ταιριάζουν στις τέσσερις κατηγορίες. Κατά δεύτερον, το πρόβλημα είναι το πώς θα εφαρμοστεί αυτό το μοντέλο σε κάθε μια από τις νέες αγορές.

**Παράδειγμα:** Έστω ότι έχουμε μια Βάση Δεδομένων πελατών μιας επιχείρησης που εμπορεύεται ηλεκτρονικούς υπολογιστές. Έστω ότι η Βάση Δεδομένων διατηρεί και της διευθύνσεις ηλεκτρονικού ταχυδρομείου των πελατών. Στις διευθύνσεις αυτές αποστέλλεται σε τακτά χρονικά διαστήματα διαφημιστικό υλικό νέων προϊόντων και εκπτώσεων. Η Βάση Δεδομένων περιέχει χαρακτηριστικά πελατών όπως ονοματεπώνυμο, ηλικία, εισόδημα, επάγγελμα και credit ratings. Οι πελάτες μπορούν να κατηγοριοποιηθούν βάσει του αν έχουν αγοράσει ή όχι ηλεκτρονικό υπολογιστή από την συγκεκριμένη εταιρία. Το να στείλουμε mail σε όλους τους πελάτες της εταιρίας (είτε έχουν αγοράσει ηλεκτρονικό υπολογιστή είτε όχι) ίσως να μην ήταν μια σωστή μια προσέγγιση. Μια καλύτερη λύση είναι το να στείλουμε mail μόνο σε όσους δεν έχουν αγοράσει ηλεκτρονικό υπολογιστή. Ένας αλγόριθμος κατηγοριοποίησης μπορεί να χρησιμοποιηθεί, ώστε να κατηγοριοποιήσει τους νέους πελάτες στις δύο κατηγορίες: πελάτες που αγόρασαν υπολογιστή, πελάτες που δεν αγόρασαν υπολογιστή.

## 1.5 ΚΑΤΗΓΟΡΙΕΣ ΑΛΓΟΡΙΘΜΩΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Όπως έχει ήδη αναφερθεί οι αλγόριθμοι κατηγοριοποίησης χωρίζονται σε δύο βασικές κατηγορίες: (i) Eager κατηγοριοποιητές και (ii) Lazy κατηγοριοποιητές. Οι Eager αλγόριθμοι κατηγοριοποίησης αρχικά κτίζουν ένα μοντέλο βασιζόμενοι στα διαθέσιμα δεδομένα εκπαίδευσης. Στη συνέχεια χρησιμοποιούν αυτό το μοντέλο για να κατηγοριοποιήσουν τα νέα στιγμιότυπα. Αντίθετα, οι Lazy αλγόριθμοι δεν κατασκευάζουν κάποιο μοντέλο. Κατηγοριοποιούν τα νέα αντικείμενα, αναλύοντας τα δεδομένα εκπαίδευσης τη στιγμή που αυτά πρέπει να κατηγοριοποιηθούν.

Ειδικότερα, μπορούμε να χωρίσουμε τους αλγόριθμους κατηγοριοποίησης σε τέσσερα είδη, τα οποία αποτελούν υποκατηγορίες των δύο παραπάνω κατηγοριών. Οι τέσσερις κατηγορίες είναι:

**Αλγόριθμοι κατηγοριοποίησης βασισμένοι σε μέτρα απόστασης:** η βασική ιδέα αυτών των αλγορίθμων είναι ότι κάθε αντικείμενο του συνόλου δεδομένων, που απεικονίζεται στην ίδια κατηγορία, θεωρείται ότι είναι πιο κοντά σε αντικείμενα της ίδιας κατηγορίας από όσο είναι σε αντικείμενα τα οποία ανήκουν σε άλλες κατηγορίες. Έτσι, μπορούν να χρησιμοποιηθούν μέτρα ομοιότητας (ή απόστασης), ώστε να οριστεί η «ομοιότητα» των διαφορετικών αντικειμένων της Βάσης Δεδομένων. Αυτή η κατηγορία αλγορίθμων ανήκει στους Lazy αλγόριθμους. Μια πολύ γνωστή και ευρέως χρησιμοποιούμενη τεχνική κατηγοριοποίησης που βασίζεται στη χρήση μέτρων απόστασης είναι αυτή των k εγγύτερων γειτόνων (k nearest neighbors – kNN) (Dasarathy, 1991). Αντίθετα, όλες οι παρακάτω κατηγορίες ανήκουν στην κατηγορία των Eager κατηγοριοποιητών.

**Αλγόριθμοι στατιστικής κατηγοριοποίησης:** δύο πολύ γνωστές μέθοδοι αυτής της κατηγορίας είναι η Bayesian κατηγοριοποίηση και η Παλινδρόμηση. Η Bayesian κατηγοριοποίηση προβλέπει τις πιθανότητες μια νέα πλειάδα να ανήκει σε μια από τις προκαθορισμένες κατηγορίες. Η απόδοση αυτού του είδους κατηγοριοποίησης είναι αρκετά υψηλή και χαρακτηρίζεται από την μεγάλη ταχύτητα της διαδικασίας κατηγοριοποίησης σε μεγάλες Βάσεις Δεδομένων. Τα προβλήματα παλινδρόμησης ασχολούνται με την εκτίμηση μιας τιμής εξόδου με βάση τις τιμές εισόδου.

**Αλγόριθμοι κατηγοριοποίησης βασισμένοι στα δένδρα απόφασης:** μια άλλη κατηγορία αλγορίθμων που χρησιμοποιούνται για την επίλυση προβλημάτων κατηγοριοποίησης είναι αυτή των Δένδρων Απόφασης (Decision Trees). Το μοντέλο κατηγοριοποίησης αυτής της κατηγορίας αλγορίθμων είναι μια δενδρική δομή. Μόλις χτιστεί η δενδρική δομή, εφαρμόζεται σε κάθε πλειάδα της Βάσης Δεδομένων και καταλήγει για κάθε μια από αυτές σε μια κατηγοριοποίηση. Η διαδικασία κατηγοριοποίησης χωρίζεται σε δύο φάσεις: (i) η κατασκευή του δένδρου και (ii) η εφαρμογή του στη Βάση Δεδομένων.

**Αλγόριθμοι κατηγοριοποίησης βασισμένοι στα Νευρωνικά Δίκτυα:** τα Νευρωνικά Δίκτυα (Neural Networks) μοντελοποιούνται με βάση τις λειτουργίες

του ανθρώπινου εγκεφάλου. Στην πραγματικότητα, τα νευρωνικά δίκτυα είναι συστήματα επεξεργασίας πληροφορίας που αποτελούνται από ένα γράφο και διάφορους αλγόριθμους που προσπελαύνουν αυτόν το γράφο. Κάθε κόμβος του γράφου είναι σαν ανεξάρτητοι νευρώνες, ενώ τα τόξα είναι σύνδεσμοι των νευρώνων. Κάθε ένας από τους κόμβους είναι στοιχείο επεξεργασίας, που λειτουργεί ανεξάρτητα από τους άλλους και χρησιμοποιεί μονό τοπικά δεδομένα που καθοδηγούν την επεξεργασία.

## ΕΠΙΛΟΓΟΣ

Η κατηγοριοποίηση είναι η πιο γνωστή και πιο δημοφιλής τεχνική εξόρυξης γνώσης. Στην εργασία αυτή θα ασχοληθούμε εκτενέστερα με το θέμα της κατηγοριοποίησης και πιο συγκεκριμένα με αλγόριθμους κατηγοριοποίησης βασισμένους σε μέτρα απόστασης.

# ΚΕΦΑΛΑΙΟ 2

# ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΕΓΓΥΤΕΡΩΝ ΓΕΙΤΟΝΩΝ

## ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό γίνεται αναφορά και περαιτέρω ανάλυση του αλγορίθμου κατηγοριοποίησης εγγύτερων γειτόνων (k-NN). Αρχικά γίνεται ανάλυση του τρόπου λειτουργίας του καθώς και κάποιων χαρακτηριστικών του, ενώ στη συνέχεια αναλύονται τα πλεονεκτήματα και μειονεκτήματα του και γίνεται αναφορά σε τεχνικές βελτίωσης του.

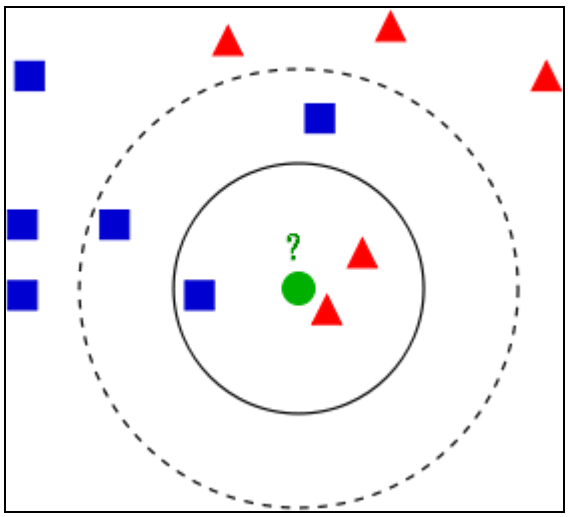
## 2.1. Ο K-NN ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ

Ένας πολύ δημοφιλής lazy κατηγοριοποιητής είναι αυτός των εγγύτερων γειτόνων (k-NN, k Nearest Neighbor) (Dasarathy, 1991). Είναι ένας instance-based αλγόριθμος. Η διαδικασία κατηγοριοποίησης προϋποθέτει απλά την αποθήκευση των δεδομένων εκπαίδευσης. Επίσης, θεωρείται ένας από τους πιο απλούς αλγόριθμους. Τα δεδομένα εκπαίδευσης τυγχάνουν επεξεργασίας όταν εμφανιστεί ένα νέο αντικείμενο. Κάθε φορά που ένα νέο αντικείμενο πρόκειται να κατηγοριοποιηθεί, υπολογίζεται η ομοιότητα του με κάθε ένα από τα αποθηκευμένα δεδομένα εκπαίδευσης.

Ως μέτρο ομοιότητας του κατηγοριοποιητή k-NN χρησιμοποιείται μια συνάρτηση απόστασης μεταξύ κάθε αντικειμένου εκπαίδευσης και του αντικειμένου που πρόκειται να κατηγοριοποιηθεί. Για την κατηγοριοποίηση ενός νέου αντικειμένου, ο κατηγοριοποιητής αναζητά και ανακτά τα k εγγύτερα αντικείμενα από το σύνολο δεδομένων εκπαίδευσης. Στη συνέχεια, του αναθέτει μια ετικέτα κατηγορίας με βάση τις κατηγορίες στις οποίες ανήκουν τα κοντινότερα αντικείμενα. Η κατηγορία αυτή είναι η πλειοψηφούσα κατηγορία που αναδεικνύεται μετά από μια διαδικασία ψηφοφορίας όπου συμμετέχουν οι k εγγύτεροι γείτονες που έχουν ανακτηθεί.

Η τιμή του k είναι ένας θετικός ακέραιος, συνήθως μικρός αριθμός και δίνεται ως παράμετρος. Στις περιπτώσεις προβλημάτων με δύο κατηγορίες, είναι προτιμότερο η τιμή του k να είναι περιττός αριθμός έτσι ώστε να αποφεύγονται οι ισοψηφίες. Για προβλήματα με περισσότερες κλάσεις, οι πιθανές ισοψηφίες επιλύονται είτε τυχαία, είτε αναθέτοντας το νέο αντικείμενο στην κατηγορία που ανήκει ο εγγύτερος γείτονας. Η βασική ιδέα του αλγορίθμου είναι ότι αν ένα αντικείμενο βρίσκεται σε μια «γειτονιά» στην οποία κυριαρχεί μια συγκεκριμένη κατηγορία, τότε αυτό σημαίνει ότι κατά πάσα πιθανότητα το αντικείμενο ανήκει σε αυτήν την συγκεκριμένη κατηγορία.

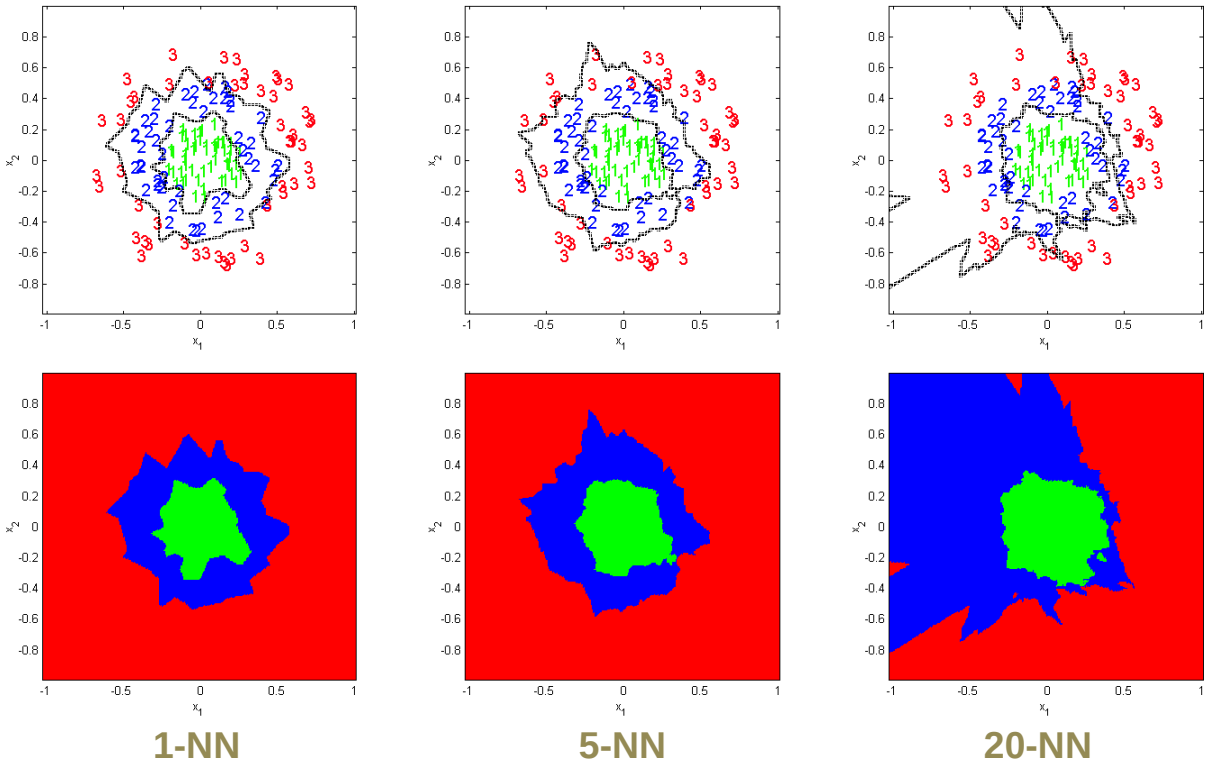
Μια πρόκληση για τον k-NN κατηγοριοποιητή είναι η επιλογή της τιμής k, καθώς διαφορετικές τιμές μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα κατηγοριοποίησης. Ένα παράδειγμα φαίνεται στο σχήμα 6. Σε αυτό, τα δεδομένα εκπαίδευσης που παρουσιάζονται είναι χωρισμένα σε δυο κατηγορίες, τα τρίγωνα και τα τετράγωνα. Ο πράσινος κύκλος είναι το αντικείμενο-ερώτημα q το οποίο πρέπει να κατηγοριοποιηθεί σε μια από τις 2 κλάσεις. Παρατηρούμε ότι για k=3, το οποίο υποδεικνύει στο σχήμα ο εσωτερικός από τους δύο ομόκεντρους κύκλους, η πρόβλεψη του αλγορίθμου είναι η κατηγορία τρίγωνο. Ενώ για k=5, το οποίο υποδεικνύει στο σχήμα ο εξωτερικός με τις διακεκομμένες γραμμές ομόκεντρος κύκλος, η πρόβλεψη είναι η κατηγορία τετράγωνο.




Σχήμα 5 "Παράδειγμα κατηγοριοποίησης k-NN αλγορίθμου"

Το k παίζει αρκετά σημαντικό ρόλο στην αποδοτικότητα του κατηγοριοποιητή και είναι δύσκολο να προσδιοριστεί. Αν η τιμή του είναι μικρή, το αποτέλεσμα μπορεί να είναι ευαίσθητο σε θορυβώδη δεδομένα. Ως θορυβώδη ορίζονται τα δεδομένα όπου τα σύνορα μεταξύ των κλάσεων δεν είναι διακριτά. Αντίθετα, αν η τιμή του k είναι πολύ μεγάλη, το αποτέλεσμα των κοντινότερων γειτόνων μπορεί να περιέχει πολλά αντικείμενα από άλλες κατηγορίες. Το k μπορεί να οριστεί χρησιμοποιώντας διάφορες ευρεστικές (heuristics) τεχνικές.

Μια ειδική περίπτωση του k-NN κατηγοριοποιητή, που χρησιμοποιείται σε πολλά ερευνητικά πεδία είναι με σταθερό k, και συγκεκριμένα με k=1. Τα μειονεκτήματα που έχει έναντι των περιπτώσεων με μεγάλο k, είναι ότι μεγάλο k σημαίνει πιο ομαλές περιοχές αποφάσεων και δίνει πιο σωστές πιθανοτικά πληροφορίες. Ωστόσο, το πολύ μεγάλο k μπορεί να χαλάσει την τοπικότητα της απόφασης και αυξάνει το υπολογιστικό κόστος. Στο σχήμα 7 είναι ένα χαρακτηριστικό παράδειγμα όπου το k=1 δίνει τα καλύτερα δυνατά αποτελέσματα. Όσο αυξάνεται το k τόσο χαλάει η τοπικότητα της απόφασης.




Σχήμα 6 "Παράδειγμα κατηγοριοποίησης k-NN αλγορίθμου για διαφορετικές τιμές του k"

Στα παραπάνω παραδείγματα έχουμε θεωρήσει ως απόσταση μεταξύ της νέας περίπτωσης και των δεδομένων εκπαίδευσης την Ευκλείδεια απόσταση. Η Ευκλείδεια απόσταση και άλλα συχνά χρησιμοποιούμενα μέτρα απόστασης σχολιάζονται στην επόμενη υποενότητα.

## 2.2. ΜΕΤΡΑ ΟΜΟΙΟΤΗΤΑΣ (ΑΠΟΣΤΑΣΗΣ)

Παρότι υπάρχουν αρκετές δυνατές επιλογές, οι περισσότεροι κατηγοριοποιητές που βασίζονται σε μέτρα ομοιότητας χρησιμοποιούν την Ευκλείδεια απόσταση. Η απόσταση ανάμεσα σε ένα νέο αντικείμενο $x' :< a_1(x'),...,a_m(x') >$ και τα αποθηκευμένα αντικείμενα εκπαίδευσης $x_i :< a_1(x_i),...,a_m(x_i) >$ (όπου m είναι ο αριθμός των χαρακτηριστικών) ορίζεται ως:

$$d(x_i, x') = \sqrt{\sum_{j=1}^{m} \left(a_j(x_i) - a_j(x')\right)^2}$$

Όταν συγκρίνουμε αποστάσεις δεν είναι απαραίτητο να υπολογίζουμε την τετραγωνική ρίζα, μπορούμε να συγκρίνουμε απευθείας τα αθροίσματα των τετραγώνων. Αυτό θα είχε ως αποτέλεσμα την επιτάχυνση της διαδικασίας

υπολογισμού της απόστασης. Μια εναλλακτική της Ευκλείδειας απόστασης είναι η απόσταση Manhattan ή city-block, όπου η διαφορά μεταξύ των τιμών των χαρακτηριστικών δεν υψώνεται στο τετράγωνο αλλά αθροίζεται (αφού έχουμε πάρει την απόλυτη τιμή της):

$$d(x_i, x') = \sum_{j=1}^{m} \left| a_j(x_i) - a_j(x') \right|$$

Άλλα μέτρα απόστασης λαμβάνονται υψώνοντας σε μεγαλύτερες δυνάμεις από το τετράγωνο. Ένα τέτοιο μέτρο είναι η απόσταση Minkowski.

$$d(x_i, x') = \left( \sum_{j=1}^{m} \left| a_j(x_i) - a_j(x') \right|^{\lambda} \right)^{1/\lambda}$$

Το λ είναι ένας ακέραιος. Αν λ=1, τότε έχουμε την απόσταση Manhattan. Αν λ=2, τότε έχουμε την απόσταση Ευκλείδεια απόσταση. Ο ρόλος του λ, όταν αυξάνεται, είναι να μεγεθύνει την απόσταση ανάμεσα στο πιο ανόμοια αντικείμενα σε σχέση με τα πιο όμοια. Γενικά, η Ευκλείδεια απόσταση αντιπροσωπεύει έναν καλό συμβιβασμό. Άλλα μέτρα απόστασης μπορεί να είναι πιο κατάλληλα σε ειδικές περιστάσεις.

Συχνά τα διαφορετικά χαρακτηριστικά μετρώνται σε διαφορετικές κλίμακες τιμών. Έτσι αν χρησιμοποιούσαμε απευθείας τον τύπο της Ευκλείδειας απόστασης, η επίδραση κάποιων από τα χαρακτηριστικά στην απόσταση δύο αντικειμένων, μπορεί να υποσκιαζόταν πλήρως από την επίδραση χαρακτηριστικών με μεγαλύτερες κλίμακες τιμών. Συνεπώς είναι συνήθης πρακτική η κανονικοποίηση των τιμών όλων των χαρακτηριστικών στο διάστημα μεταξύ 0 και 1.

Μέχρι τώρα έχουμε υποθέσει την ύπαρξη αριθμητικών χαρακτηριστικών. Στα αριθμητικά χαρακτηριστικά, η διαφορά μεταξύ δύο τιμών είναι απλά η αριθμητική τους διαφορά. Για διακριτά χαρακτηριστικά τα οποία λαμβάνουν συμβολικές και όχι αριθμητικές τιμές, συνηθίζεται να παίρνουμε ως 1 την απόσταση μεταξύ δύο τιμών που δεν ταυτίζονται και ως 0 την απόσταση όταν οι τιμές ταυτίζονται. Σε αυτήν την περίπτωση δεν απαιτείται κανονικοποίηση αφού χρησιμοποιούνται μόνο οι τιμές 0 και 1.

Τέλος θα πρέπει να αναφέρουμε ότι υπάρχουν και άλλες, λιγότερο γνωστές αποστάσεις που έχουν εμφανιστεί στην βιβλιογραφία. Συγκεκριμένα, έχουν προταθεί ως μέτρα ομοιότητας η απόσταση Canberra, ο συντελεστής Czekanowski, απόσταση Chebychev ή Maximum. Θα ήταν λάθος να θεωρήσουμε ότι ως μέτρα ομοιότητας χρησιμοποιούνται μόνο οι διάφοροι μαθηματικοί τύποι αποστάσεων. Υπάρχουν και άλλα μέτρα ομοιότητας που έχουν εφαρμοστεί σε συστήματα ανάκτησης πληροφορίας και στις μηχανές αναζήτησης στο διαδίκτυο. Μερικά από τα μέτρα αυτά είναι το μέτρο Dice, Jaccard, το συνημίτονο και η επικάλυψη.

## 2.3. ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΟΥ k-NN ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗ

Ο kNN κατηγοριοποιητής συγκαταλέγεται ανάμεσα στους δημοφιλέστερους του είδους του. Τα πλεονεκτήματα που έχει και τον καθιστούν τόσο δημοφιλή είναι τα παρακάτω:

- Η λειτουργία του είναι εύκολο να κατανοηθεί από τον άνθρωπο
- Είναι απλός στην υλοποίηση. Απαιτείται μόνο μια ακέραια τιμή για την παράμετρο k, ένα σύνολο εκπαίδευσης και ένα μέτρο απόστασης.
- Έχει πολλές εφαρμογές όπως αναγνώριση προτύπων, κατηγοριοποίηση χρονοσειρών κτλ.
- Δουλεύει καλά σε περιπτώσεις multi-modal κλάσεων και σε εφαρμογές όπου κάποιο αντικείμενο μπορεί να ανήκει σε περισσότερο από μια κλάσης.
- Είναι αποδοτικός σε περιπτώσεις όπου τα δεδομένα εκπαίδευσης περιέχουν θόρυβο (noisy) και σε περιπτώσεις όπου τα δεδομένα εκπαίδευσης είναι πολλά ($N \rightarrow \infty$).
- Δε χρειάζεται να σχεδιάσει ή να μάθει κάποιο μοντέλο και έτσι μπορεί να προσαρμόζεται εύκολα σε αλλαγές στα δεδομένα εκπαίδευσης.

Το τελευταίο πλεονέκτημα πηγάζει από το γεγονός ότι είναι ένας lazy αλγόριθμος, δηλαδή δεν γίνεται οποιαδήποτε εκπαίδευση μέχρι να φτάσει κάποιο αντικείμενο για κατηγοριοποίηση. Παράλληλα όμως, το πλεονέκτημα αυτό είναι και η αιτία του πρώτου από τα παρακάτω μειονεκτήματα του κατηγοριοποιητή το οποίο αποτελεί και το κύριο αντικείμενο αυτής της εργασίας:

- Το κόστος υπολογισμού όλων των αποστάσεων είναι μεγάλο. Ο αλγόριθμος πρέπει να υπολογίσει την απόσταση κάθε αντικειμένου προς κατηγοριοποίηση με κάθε αντικείμενο τους συνόλου εκπαίδευσης. Έτσι για παράδειγμα αν έχουμε ένα σύνολο δεδομένων 15.000 αντικειμένων, όπου τα 10.000 είναι τα δεδομένα εκπαίδευσης και τα 5.000 είναι τα νέα αντικείμενα προς κατηγοριοποίηση, οι υπολογισμοί που πρέπει να γίνουν είναι 10.000 * 5.000 = 50.000.000 υπολογισμοί αποστάσεων. Όσο μεγαλώνει η βάση δεδομένων το κόστος των υπολογισμών αυτών αυξάνεται εκθετικά. Επιπρόσθετα, το κόστος κάθε υπολογισμού απόστασης εξαρτάται από το πλήθος των χαρακτηριστικών (attributes) των αντικειμένων. Για παράδειγμα. σε περιπτώσεις δεδομένων χρονοσειρών, όπου κάθε αντικείμενο περιγράφεται από έναν μεγάλο αριθμό χαρακτηριστικών, το κόστος αυτό είναι ακόμη μεγαλύτερο.
- Ένα ακόμα μειονέκτημα είναι ότι έχει μεγάλη απαίτηση σε αποθηκευτικό χώρο. Σε αντίθεση με τους eager κατηγοριοποιητές, οι οποίοι αρχικά κατασκευάζουν ένα μοντέλο κατηγοριοποίησης και στην συνέχεια μπορούν να διαγράψουν τα δεδομένα εκπαίδευσης, ο αλγόριθμος k εγγύτερων γειτόνων προϋποθέτει ότι τα δεδομένα εκπαίδευσης είναι πάντα στην

διάθεση του. Έτσι, σε καμία περίπτωση δεν μπορούν να διαγραφούν από τη μνήμη.

## 2.4. ΠΟΛΥΔΙΑΣΤΑΤΑ ΔΕΔΟΜΕΝΑ ΚΑΙ k-NN ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ

Ο αλγόριθμος k-NN υπολογίζει την απόσταση μεταξύ δύο αντικειμένων με βάση όλα τα χαρακτηριστικά. Αυτό μπορεί να μειώσει σημαντικά την ακρίβεια του αλγορίθμου όταν υπάρχουν πολλά χαρακτηριστικά που δεν επηρεάζουν την κατηγορία. Έστω ένα πρόβλημα με 30 χαρακτηριστικά, εκ των οποίων μόνο τα 5 είναι σημαντικά για την πρόβλεψη νέων περιπτώσεων. Τότε στην περίπτωση που ένα αντικείμενο έχει τις ίδιες τιμές στα 5 αυτά χαρακτηριστικά, αλλά διαφορετικές σε όλα τα άλλα, μπορεί να έχει πολύ μεγάλη Ευκλείδεια απόσταση και κατά συνέπεια να κατηγοριοποιηθεί λανθασμένα.

Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν προταθεί μέθοδοι τόσο για τη στάθμιση των χαρακτηριστικών, όσο και για την επιλογή ενός υποσυνόλου χαρακτηριστικών. Και στις δύο περιπτώσεις απαιτείται η εύρεση των σημαντικότερων χαρακτηριστικών για την πρόβλεψη της κατηγορίας. Συχνά κάποια χαρακτηριστικά είναι περισσότερο σημαντικά όσον αφορά μία από τις τιμές μίας διακριτής κατηγορίας και λιγότερο σημαντικά σε σχέση με κάποια άλλη τιμή. Σε αυτήν την περίπτωση απαιτείται ένας διαχωρισμός μεταξύ των σημαντικών χαρακτηριστικών για κάθε διακριτή τιμή της κλάσης. Υπάρχει ένας τομέας της Μηχανικής Μάθησης που ασχολείται με τα παραπάνω ζητήματα ονομάζεται Επιλογή Χαρακτηριστικών (Feature Selection).

## 2.5. ΧΡΟΝΟΣΕΙΡΕΣ ΚΑΙ k-NN ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ

Με τον όρο χρονοσειρά εννοούμε μια ακολουθία $\{x_t : t=0,1,2,\ldots\}$ όπου κάθε $x_t$ εκφράζει την κατάσταση κατά την χρονική στιγμή *t,* ενός συστήματος το οποίο εξελίσσεται στο χρόνο. Παραδείγματα τέτοιων χρονοσειρών είναι:

- Οι ημερήσιες, αεροπορικές και οδικές, αφίξεις τουριστών στην χώρα μας
- Ο αριθμός πελατών μέσα σε ένα πολυκατάστημα κατά τη διάρκεια της μέρας
- Η ημερήσια κατανάλωση ηλεκτρικού ρεύματος καθώς και η ημερήσια κατανάλωση ύδατος
- Οι οικονομικές χρονοσειρές, όπως η καθημερινή κίνηση των τιμών μιας μετοχής στο χρηματιστήριο
- Οι μετεωρολογικές χρονοσειρές, όπως η θερμοκρασία περιβάλλοντος και ατμοσφαιρική πίεση σε συγκεκριμένες χρονικές στιγμές.

Στην πραγματικότητα, τα δεδομένα χρονοσειρών είναι σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών, κάθε ένα από τα οποία απεικονίζει την τιμή μιας

παρατήρησης ενός αντικειμένου μια συγκεκριμένη χρονική στιγμή. Ο αλγόριθμος k εγγύτερων γειτόνων έχει εφαρμοστεί με επιτυχία σε προβλήματα κατηγοριοποίησης χρονοσειρών. Η τιμή του k η οποία ενδείκνυται σε αυτές τις περιπτώσεις δεδομένων είναι το k=1. Στο πειραματικό μέρος αυτής της πτυχιακής έγιναν δοκιμές και για άλλες τιμές του k, αλλά η προαναφερθείσα τιμή έδινε πάντα μεγαλύτερο ποσοστό ακρίβειας. Ασφαλώς, το πρόβλημα που υφίσταται σε αυτού του είδους τα δεδομένα, είναι το αυξημένο κόστος υπολογισμού απόστασης δύο σημείων εξαιτίας του μεγάλου πλήθους των παρατηρήσεων (χαρακτηριστικών). Για την μείωση του κόστους, συχνά εφαρμόζονται τεχνικές μείωσης των διαστάσεων (dimensionality reduction).

## 2.6. ΤΕΧΝΙΚΕΣ ΕΠΙΤΑΧΥΝΣΗΣ ΤΟΥ k-NN ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗ

Ο κατηγοριοποιητής k-NN χρησιμοποιείται σε πολλές εφαρμογές χάρη στην απλότητα και αποτελεσματικότητα του. Ο απλούστερος αλγόριθμος για να υλοποιηθεί ο κατηγοριοποιητής είναι η σειριακή αναζήτηση. Αυτό σημαίνει ότι για κάθε αντικείμενο x προς κατηγοριοποίηση, ο κατηγοριοποιητής k-NN θα υπολογίσει όλες τις αποστάσεις μεταξύ του x και όλων των δεδομένων εκπαίδευσης. Αυτό, όπως έχει ήδη αναφερθεί, είναι το σημαντικότερο μειονέκτημα του k-NN κατηγοριοποιητή και αποτελεί ένα ανοιχτό πεδίο έρευνας που έχει προσελκύσει το ενδιαφέρον ερευνητών που προέρχονται από διαφορετικές ερευνητικές περιοχές της πληροφορικής, όπως οι Βάσεις Δεδομένων, η Μηχανική Μάθηση και Τεχνίτη νοημοσύνη, η Στατιστική και η εξόρυξη δεδομένων. Αποτέλεσμα των ερευνητικών προσπαθειών τους είναι η δημοσίευση πολλών εργασιών που προτείνουν μεθόδους για την αντιμετώπιση αυτού του σημαντικού μειονεκτήματος του k-NN κατηγοριοποιητή. Οι μέθοδοι που έχουν προταθεί μπορούν να κατηγοριοποιηθούν σε δύο βασικές κατηγορίες επιτάχυνσης του k-NN κατηγοριοποιητή: (i) Μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων (Multiatribute indexing) και (ii) Μέθοδοι μείωσης του όγκου των δεδομένων (Data Reduction Techniques).

### 2.6.1. ΜΕΘΟΔΟΙ ΔΕΙΚΤΟΔΟΤΗΣΗΣ ΔΕΔΟΜΕΝΩΝ (INDEXING METHODS)

Οι μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων (Zezula et al, 2006), (Samet, 2006) αποτελούν τα αποτελέσματα έρευνας επιστημόνων που προέρχονται από την περιοχή των Βάσεων Δεδομένων. Οι μέθοδοι δεικτοδότησης απαιτούν την προ-επεξεργασία των διαθέσιμων δεδομένων εκπαίδευσης ώστε να κατασκευαστεί η δομή του δείκτη, η οποία συνήθως έχει δενδροειδή μορφή. Οι μέθοδοι αναζήτησης κοντινότερων γειτόνων σε τέτοιου είδους δομές είναι συνήθως πολύ αποτελεσματικές, λόγω του ότι είναι ικανές να αποφύγουν πολλούς υπολογισμούς αποστάσεων.

Έχουν προταθεί δεκάδες μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων. Χαρακτηρίστηκα παραδείγματα είναι το R-tree (Gutman, 1984) και οι διάφορες παραλλαγές του (Manolopoulos, 2006), το k-DB-tree (Robinson, 1981) και το Vantage Point (VP) tree (Yanilos, 1993). Το βασικό μειονέκτημα των περισσότερων μεθόδων δεικτοδότησης είναι ότι η απόδοση της αναζήτησης εγγύτερων γειτόνων σε δεικτοδοτημένα δεδομένα εξαρτάται σε μεγάλο βαθμό από το πλήθος των διαστάσεων. Η απόδοση (ή ταχύτητα εκτέλεσης) είναι σε υψηλά επίπεδα όταν τα δεδομένα που χρησιμοποιούμε είναι λίγων διαστάσεων (συνήθως έως 10). Σε υψηλότερες διαστάσεις η απόδοση τους σταδιακά μειώνεται και μπορεί να φθάσει ακόμη και στα επίπεδα (ή και χειρότερη) της σειριακής αναζήτησης. Το φαινόμενο αυτό είναι γνωστό ως φαινόμενο της κατάρας των διαστάσεων (dimensionality curse). Συνεπώς, οι μέθοδοι δεικτοδότησης μπορούν να επιταχύνουν την αναζήτηση εγγύτερων γειτόνων όταν τα δεδομένα που έχουμε στην διάθεση μας είναι λίγων διαστάσεων.

Όπως έχουμε ήδη αναφέρει υπάρχουν τεχνικές μείωσης των διαστάσεων, όπως αυτή της ανάλυσης κύριων συνιστωσών (Principal Component Analysis - PCA) (Jolliffe, 2002). Το πρόβλημα δεικτοδότησης δεδομένων με μεγάλο αριθμό διαστάσεων μπορεί να αντιμετωπιστεί χρησιμοποιώντας μια τέτοια μέθοδο. Ωστόσο, η εφαρμογή της αποτελεί ένα επιπρόσθετο βήμα προ-επεξεργασίας η οποία εισάγει επιπρόσθετο κόστος. Επίσης, δεν είναι πάντα επιτυχείς και μπορεί να ισοδυναμεί με χάσιμο χρήσιμης πληροφορίας. Τέλος, για την κατηγοριοποίηση κάθε νέου στιγμιότυπου απαιτείται ο μετασχηματισμός του ώστε να είναι ίδιας διάστασης με τα μετασχηματισμένα δεδομένα εκπαίδευσης.

Είναι σημαντικό να αναφερθεί, ότι αν και οι μέθοδοι δεικτοδότησης πολυδιάστατων δεδομένων μπορούν να επιταχύνουν την αναζήτηση εγγύτερων γειτόνων, σε αντίθεση με τις τεχνικές μείωσης του όγκου των δεδομένων, δεν μειώνουν τις απαιτήσεις χώρου για την αποθήκευση των δεδομένων. Έτσι, αυτές οι μέθοδοι και κατ' επέκταση ο κατηγοριοποιητής k-NN, δεν μπορούν να εφαρμοσθούν σε συσκευές με περιορισμένη μνήμη.

### 2.6.2. ΤΕΧΝΙΚΕΣ ΜΕΙΩΣΗΣ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ (DATA REDUCTION TECHNIQUES)

Οι τεχνικές μείωσης του όγκου των δεδομένων χωρίζονται σε δυο μεγάλες κατηγορίες: (i) τεχνικές μείωσης των δεδομένων εκπαίδευσης και (ii) τεχνικές μείωσης των διαστάσεων.

Οι τεχνικές μείωσης των δεδομένων εκπαίδευσης έχουν στόχο την γρήγορη κατηγοριοποίηση με βάση την μέθοδο εγγύτερων γειτόνων. Αποτελούν προσπάθειες ερευνητών που προέρχονται από τον χώρο της Μηχανικής Μάθησης και έχουν σαν στόχο την κατασκευή ενός μικρού συνόλου δεδομένων εκπαίδευσης, το οποίο αντιπροσωπεύει όσο το δυνατό περισσότερο το αρχικό, μεγάλο σε όγκο σύνολο. Το μικρό αυτό σύνολο αποτελείται από λίγα και

αντιπροσωπευτικά αντικείμενα εκπαίδευσης. Η σειριακή αναζήτηση εγγύτερων γειτόνων σε αυτό το μικρό, αντιπροσωπευτικό σύνολο, σε αντίθεση με την εφαρμογή της στο αρχικό σύνολο, μπορεί να εφαρμοστεί χωρίς την σπατάλη υψηλού υπολογιστικού κόστους.

Αντίθετα, οι τεχνικές μείωσης διαστάσεων έχουν στόχο στην μείωση του κόστους υπολογισμού αποστάσεων. Όπως έχει αναφερθεί, το κόστος υπολογισμού απόστασης εξαρτάται από το πλήθος των χαρακτηριστικών των αντικειμένων. Η μείωση τους ισοδυναμεί με μείωση του κόστους, δηλαδή με την επιτάχυνση την διαδικασία κατηγοριοποίησης.

Τα επόμενα κεφάλαια αυτής της εργασίας είναι αφιερωμένα αποκλειστικά στις τεχνικές μείωσης όγκου και έτσι δεν επεκτεινόμαστε περισσότερο σε αυτό το σημείο της εργασίας.

## ΕΠΙΛΟΓΟΣ

Είναι γεγονός ότι ο κατηγοριοποιητής εγγύτερων γειτόνων είναι πολύ δημοφιλής και χρησιμοποιείται σε πολλές εφαρμογές χάρη στην απλότητα και αποτελεσματικότητα του. Το αυξημένο μέγεθος των βάσεων δεδομένων που καλείται να επεξεργαστεί τις περισσότερες φορές, αλλά και η ταχύτητα η οποία θέλουμε λάβουμε τα αποτελέσματα για την περαιτέρω επεξεργασία τους, καθιστά σχεδόν απαραίτητη τη χρήση ορισμένων τεχνικών επιτάχυνσης του. Στο επόμενο κεφάλαιο θα ασχοληθούμε με μεθόδους μείωσης του όγκου δεδομένων για να επιτευχθεί ο παραπάνω σκοπός.

# ΚΕΦΑΛΑΙΟ 3

# ΤΕΧΝΙΚΕΣ ΜΕΙΩΣΗΣ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ

## ΕΙΣΑΓΩΓΗ

Αυτό το κεφάλαιο της εργασίας αφιερώνεται αποκλειστικά στις μεθόδους μείωσης όγκου δεδομένων. Συγκεκριμένα, στις τεχνικές μείωσης δεδομένων εκπαίδευσης (ΤΜΔΕ) και σε αυτές της μείωσης του πλήθος των διαστάσεων-χαρακτηριστικών (ΤΜΠΔ). Απώτερος σκοπός είναι η επιτάχυνση του κατηγοριοποιητή εγγύτερων γειτόνων.

## 3.1. ΤΕΧΝΙΚΕΣ ΜΕΙΩΣΗΣ ΔΕΔΟΜΕΝΩΝ ΕΚΠΑΙΔΕΥΣΗΣ (ΤΜΔΕ)

### 3.1.1 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Ένας αποτελεσματικός τρόπος για την επιτάχυνση του k-NN κατηγοριοποιητή είναι η μείωση των δεδομένων εκπαίδευσης, ώστε η σειριακή αναζήτηση των εγγύτερων γειτόνων να επιτυγχάνεται χωρίς ιδιαίτερο κόστος. Οι ΤΜΔΕ έχουν σαν στόχο την κατασκευή ενός μικρού συνόλου δεδομένων το οποίο αντιπροσωπεύει όσο το δυνατό περισσότερο το αρχικό, μεγάλο σε όγκο σύνολο. Το σύνολο αυτό συνήθως ονομάζεται συμπυκνωμένο σύνολο (condensing set). Με άλλα λόγια, στόχος είναι η μείωση του υπολογιστικού κόστους αναζήτησης και ταυτόχρονα διατήρηση της ακρίβειας κατηγοριοποίησης σε υψηλά επίπεδα. Η μεθοδολογία που υιοθετούν οι ΤΜΔΕ, επιπρόσθετα του πλεονεκτήματος της γρήγορης κατηγοριοποίησης νέων αντικειμένων, έχει το πλεονέκτημα της μείωσης ανάγκης για μεγάλη μνήμη, σε αντίθεση με τις μεθόδους δεικτοδότησης. Έτσι, η κατηγοριοποίηση νέων αντικειμένων μπορεί να εκτελείται από συσκευές με περιορισμένη μνήμη (π.χ. αισθητήρες που λαμβάνουν δεδομένα από το περιβάλλον τους) και χωρίς την ανάγκη της μεταφοράς τους σε ισχυρούς υπολογιστές, πράγμα που προϋποθέτει συνδεσιμότητα και πολλές φορές ισοδυναμεί με υψηλό κόστος.

Οι ΤΜΔΕ μπορούν να διακριθούν σε δύο μεγάλες κατηγορίες αλγορίθμων: (i) αλγόριθμοι επιλογής δεδομένων (data selection algorithms) και (ii) αλγόριθμοι σύνοψης (ή παραγωγής) δεδομένων (data abstraction (or generation) algorithms). Και οι δύο κατηγορίες έχουν στόχο τη δημιουργία ενός μικρού αντιπροσωπευτικού συνόλου δεδομένων, αλλά διαφέρουν στην μεθοδολογία που χρησιμοποιούν για να το κατασκευάσουν.

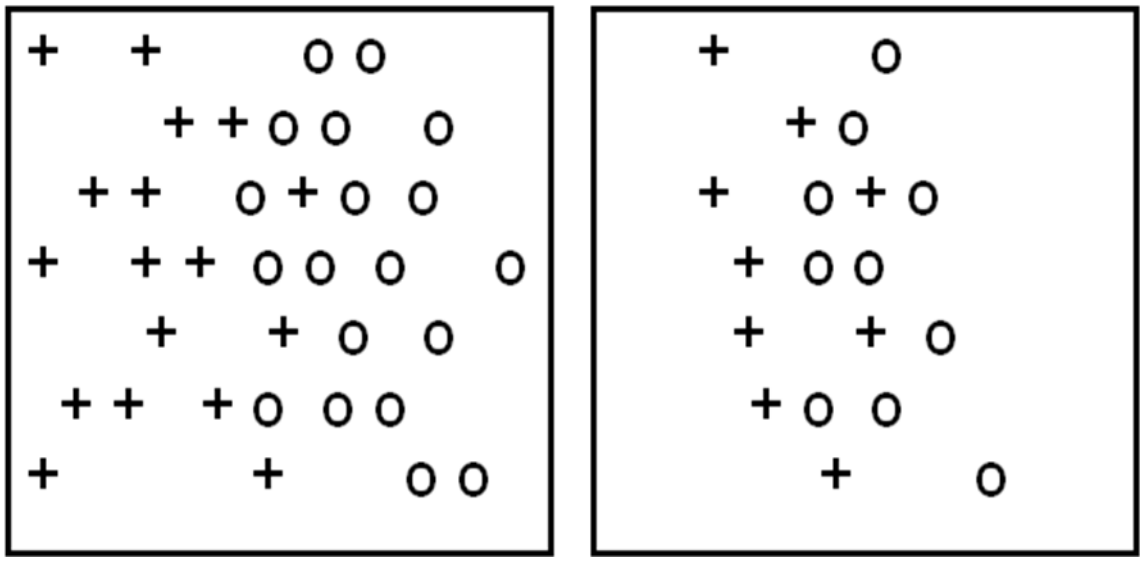
Οι αλγόριθμοι επιλογής δεδομένων (Garcia et al, 2011) επιλέγουν κάποια στιγμιότυπα από το αρχικό σύνολο δεδομένο εκπαίδευσης ως αντιπροσώπους και

τα αποθηκεύουν στο συμπυκνωμένο σύνολο. Αντίθετα, οι αλγόριθμοι σύνοψης (Triguero et al, 2012) δημιουργούν νέους αντιπροσώπους συνοψίζοντας όμοια στιγμιότυπα του αρχικού συνόλου και τα τοποθετούν στο συμπυκνωμένο σύνολο (π.χ. βρίσκοντας το μέσο όρο των τιμών των χαρακτηριστικών, εφαρμόζοντας αλγορίθμους ομαδοποίησης (clustering) και συγχώνευσης όμοιων στιγμιότυπων). Οι πρόσφατες εργασίες (Garcia et al, 2011) και (Triguero et al 2012) παρουσιάζουν ενδιαφέρουσες ταξινομήσεις των αλγορίθμων επιλογής και σύνοψης δεδομένων αντίστοιχα. Επίσης, στις εργασίες αυτές οι αντίστοιχοι αλγόριθμοι παρουσιάζονται και συγκρίνονται μεταξύ τους.

Μια άλλη κατηγοριοποίηση των μεθόδων μείωσης όγκου που εμφανίζεται στη βιβλιογραφία είναι εξής: (i) επαυξητικοί (incremental) αλγόριθμοι και (ii) μη επαυξητικοί (decremental) αλγόριθμοι. Οι επαυξητικοί αλγόριθμοι ξεκινούν έχοντας ένα ελάχιστο όγκο δεδομένων, που σταδιακά τον αυξάνουν μέχρι να ικανοποιήσουν τα κριτήρια που θέτει ο αλγόριθμος. Αντίθετα, οι μη επαυξητικοί αλγόριθμοι αρχικά λαμβάνουν υπόψη όλα τα δεδομένα εκπαίδευσης και σταδιακά τα μειώνουν μέχρι να ικανοποιηθούν τα κριτήρια.

Οι περισσότερες ΤΜΔΕ προσπαθούν να εντοπίσουν τα «σύνορα» μεταξύ των διαφορετικών κλάσεων στα δεδομένα. Στο συμπυκνωμένο σύνολο τοποθετούνται περισσότερα αντικείμενα από τις περιοχές που βρίσκονται κοντά σε σύνορα κλάσεων παρά από τις «εσωτερικές» περιοχές αυτών. Η λειτουργία αυτή βασίζεται στην ιδέα του ότι τα δεδομένα που δεν ορίζουν τα όρια απόφασης, δεν επηρεάζουν την πρόβλεψη κατηγοριοποίησης και έτσι μπορούν να απομακρυνθούν. Αντίθετα, μόνο τα στιγμιότυπα που ορίζουν τα όρια απόφασης επηρεάζουν το αποτέλεσμα της κατηγοριοποίησης και έτσι πρέπει να τοποθετηθούν στο συμπυκνωμένο σύνολο.

Η προαναφερθείσα ιδέα γίνεται ευκολότερα κατανοητή παρατηρώντας το σχήμα 7. Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων εκπαίδευσης με δύο κλάσεις. Συγκεκριμένα, έχουμε τα στιγμιότυπα που ανήκουν στην κλάση «σταυρός» και αυτά που ανήκουν στην κλάση «κύκλος». Στην αριστερή πλευρά του σχήματος παρουσιάζεται το αρχικό σύνολο, ενώ στα δεξιά το συμπυκνωμένο που προέκυψε μετά την εφαρμογή ενός αλγορίθμου μείωσης όγκου δεδομένων. Ένα νέο, μη κατηγοριοποιημένο αντικείμενο θα κατηγοριοποιηθεί στην κλάση «σταυρός» αν βρεθεί από την αριστερή πλευρά του «συνόρου» απόφασης, είτε χρησιμοποιηθεί το αρχικό είτε το συμπυκνωμένο σύνολο δεδομένων. Αντίθετα, αν βρεθεί στην δεξιά πλευρά θα κατηγοριοποιηθεί στην κλάση «κύκλος».

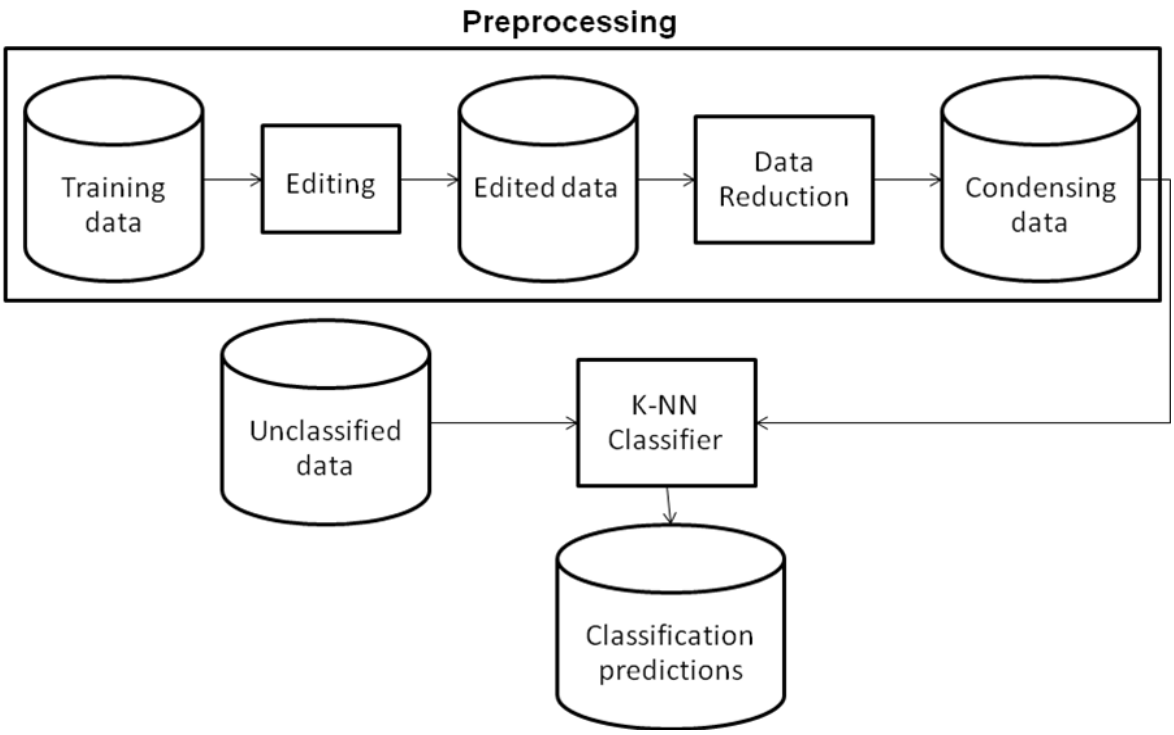


Σχήμα 7 "Παράδειγμα μείωσης όγκου δεδομένων (αριστερά: Αρχικό σύνολο, Δεξιά: Συμπυκνωμένο σύνολο)"

Τέλος, είναι σημαντικό να αναφερθεί ότι μια υποκατηγορία των αλγορίθμων επιλογής έχουν σαν βασικό στόχο την αύξηση της ακρίβειας κατηγοριοποίησης παρά τη μείωση του κόστους. Οι αλγόριθμοι που ανήκουν σε αυτή την υποκατηγορία καλούνται αλγόριθμοι επεξεργασίας δεδομένων (editing algorithms). Οι αλγόριθμοι αυτοί, προσπαθούν να αυξήσουν την ακρίβεια απομακρύνοντας τα δεδομένα με θόρυβο (noisy items), καθώς επίσης «ξεχωρίζοντας» ή «εξομαλύνοντας» τα όρια απόφασης των κλάσεων.

Ο όγκος μείωσης δεδομένων εκπαίδευσης εξαρτάται από το πλήθος των κλάσεων καθώς επίσης και από το επίπεδο θορύβου που υπάρχει στα δεδομένα. Όσο περισσότερος θόρυβος υπάρχει στα δεδομένα, τόσο περισσότερα δεδομένα αποθηκεύονται στο συμπυκνωμένο σύνολό. Επίσης, όσες περισσότερες κλάσεις υπάρχουν, τόσα περισσότερα όρια απόφασης υπάρχουν, γεγονός που συνεπάγεται με μεγαλύτερο συμπυκνωμένο σύνολό. Έτσι, πολλές φορές, η αποδοτική εφαρμογή μιας ΤΜΔΕ προϋποθέτει το να έχει απομακρυνθεί ο θόρυβος από τα δεδομένα από έναν αλγόριθμο επεξεργασίας (Lozano, 2007).

Γίνεται αντιληπτό λοιπόν, ότι συχνά η διαδικασία κατηγοριοποίησης αποτελεί μια ακολουθία εφαρμογής αλγόριθμων. Αρχικά, στο σύνολο δεδομένων εκπαίδευσης εφαρμόζεται ένας αλγόριθμος επεξεργασίας ώστε να απομακρυνθεί ο θόρυβος. Το σύνολο δεδομένων που προκύπτει ονομάζεται επεξεργασμένο σύνολο (edited data). Στην συνέχεια σε αυτό το σύνολο εφαρμόζεται μια ΤΜΔΕ. Το νέο σύνολο που προκύπτει ονομάζεται συμπυκνωμένο σύνολο δεδομένων. Η προαναφερθείσα διαδικασία αποτελεί την προ-επεξεργασία των δεδομένων εκπαίδευσης. Για κάθε νέο αντικείμενο, ο k-NN κατηγοριοποιητής εφαρμόζεται στο συμπυκνωμένο σύνολο δεδομένων εκπαίδευσης και προβλέπει την κλάση του νέου αντικειμένου. Η προαναφερθείσα διαδικασία συνοψίζεται στο σχήμα 8.




**Σχήμα 8 "Διαδικασία απομάκρυνσης θορύβου, μείωσης όγκου και εφαρμογής του k-NN κατηγοριοποιητή"**

Εκτός από τις πρόσφατες εργασίες (Garcia et al, 2011) και (Triguero et al 2012), άλλες εργασίες ανασκόπησης της βιβλιογραφίας σχετικά με τις ΤΜΔΕ είναι διαθέσιμες στις αναφορές: (Lozano, 2007), (Toussaint, 2002), (Wilson και Martinez, 2000), (Jankowski και Grochowski, 2004), (Grochowski και Jankowski, 2004), (Lopez et al, 2010), (Brighton, 2004).

Στη συνέχεια του κεφαλαίου αυτού παρουσιάζονται αναλυτικά τρεις γνωστές ΤΜΔΕ:

- αλγόριθμος επιλογής CNN-rule,
- αλγόριθμος σύνοψης ο RSP3 και
- αλγόριθμος επεξεργασίας Wilson editing

Παράλληλα, είτε παρουσιάζονται συνοπτικά είτε αναφέρονται άλλες ΤΜΔΕ που προτείνονται στη σχετική βιβλιογραφία. Στην συνέχεια της εργασίας εκτιμάτε η απόδοση των τριών τεχνικών και συγκρίνεται εφαρμόζοντας την ακολουθία διαδικασιών του σχήματος 8.

### 3.1.2 ΚΑΝΟΝΑΣ ΣΥΜΠΥΚΝΩΣΗΣ ΕΓΓΥΤΕΡΟΥ ΓΕΙΤΟΝΑ (CNN-rule)

Ο Hart, το 1968, ήταν ο πρώτος που παρουσίασε την έννοια της μείωσης του όγκου των δεδομένων εκπαίδευσης με στόχο την γρήγορη εφαρμογή του k-NN κατηγοριοποιητή. Επίσης, παρουσίασε για πρώτη φόρα το γεγονός, ότι τα αντικείμενα του συνόλου εκπαίδευσης που δε βρίσκονται κοντά στα σύνορα απόφασης των κλάσεων μπορούν να απομακρυνθούν με ασφάλεια, με αποτέλεσμα το κόστος της σειριακής αναζήτησης των γειτόνων να μειωθεί σε μεγάλο βαθμό. Με αυτόν τον τρόπο ο Hart εισήγαγε τον πρώτο, ένα από τους πιο γνωστούς και με μεγάλη συχνότητα χρήσης, αλγορίθμους μείωσης όγκου δεδομένων εκπαίδευσης (Hart, 1968). Ο αλγόριθμος αυτός είναι γνωστός στη βιβλιογραφία με το όνομα κανόνας συμπύκνωσης εγγύτερου γείτονα (Condensing Nearest Neighbor rule – CNN-rule). Ο αλγόριθμος του Hart έχει υιοθετηθεί ως μέτρο σύγκρισης (σημείο αναφοράς) στις περισσότερες εργασίες που παρουσιάζουν νέες ΤΜΔΕ. Ακόμα και σήμερα αποτελεί έναν από τους πιο αποδοτικούς αλγορίθμους. Επίσης, πολλοί μεταγενέστεροι αλγόριθμοι αποτελούν παραλλαγές ή έχουν βασιστεί στην ιδέα του CNN-rule. Παράδειγμα τέτοιων αλγορίθμων είναι:

- ο κανόνας μείωσης εγγύτερου γείτονα (Reduced Nearest Neighbor Rule) (Gates, 1972)
- ο κανόνας επιλογής εγγύτερου γείτονα (Selective Nearest Neighbor Rule) (Ritter, 1975)
- ο τροποποιημένος κανόνας συμπύκνωσης εγγύτερου γείτονα (Modified Nearest Neighbor Rule) (Devi και Murty, 2002)
- ο γρήγορος κανόνας συμπύκνωσης εγγύτερου γείτονα (Fast Nearest Neighbor Rule) (Devi και Murty, 2002)
- Οι οικογένεια αλγόριθμων IB (Aha et al, 1991)

Ο αλγόριθμος του Hart είναι ένας αλγόριθμος επιλογής που προσπαθεί να αποθηκεύσει στο συμπυκνωμένο σύνολο εκπαίδευσης μόνο τα αντικείμενα που βρίσκονται κοντά στα σύνορα απόφασης. Αυτό επιτυγχάνεται χρησιμοποιώντας δύο σύνολα δεδομένων, έστω Α και Β. Αρχικά το πρώτο αντικείμενο του συνόλου δεδομένων εκπαίδευσης αποθηκεύεται στο σύνολο Α και όλα τα υπόλοιπα αντικείμενα στο σύνολο Β. Στην συνέχεια, ο CNN-rule προσπαθεί να κατηγοριοποιήσει τα περιεχόμενα του συνόλου Β αναζητώντας τον εγγύτερο γείτονα (1-NN) στο σύνολο Α. Όποιο αντικείμενο του Β δεν κατηγοριοποιείται σωστά, μεταφέρεται από το ένα σύνολο στο άλλο, δηλαδή από το Β στο Α. Αντίθετα, τα αντικείμενα που κατηγοριοποιούνται σωστά δεν μεταφέρονται. Η διαδικασία αυτή συνεχίζει μέχρι να σταματήσουν οι μεταφορές από το Β στο Α, δηλαδή όλα τα στιγμιότυπα που ανήκουν στο Β να κατηγοριοποιούνται σωστά εφαρμόζοντας τον 1-NN κατηγοριοποιητή στα δεδομένα του Α. Το τελικό

συμπυκνωμένο σύνολο εκπαίδευσης είναι το σύνολο δεδομένων Α. Η διαδικασία αυτή συνοψίζεται στον αλγόριθμο που παρουσιάζεται στο σχήμα 9.

Η ιδέα στην οποία βασίζεται ο αλγόριθμος είναι η εξής: Αν η κλάση ενός αντικειμένου διαφωνεί με την κλάση ενός γειτονικού αντικειμένου, τότε βρίσκεται κοντά στα σύνορα απόφασης και έτσι πρέπει να συμπεριληφθεί στο συμπυκνωμένο σύνολο εκπαίδευσης. Σε διαφορετική περίπτωση, το αντικείμενο βρίσκεται σε εσωτερική περιοχή της κλάσης (βλέπε σχήμα 7) και έτσι δεν συμπεριλαμβάνεται στο συμπυκνωμένο σύνολο εκπαίδευσης.

Ένα από τα πλεονεκτήματα του CNN-rule είναι το ότι καθορίζει το μέγεθος του συμπυκνωμένου συνόλου δεδομένων αυτόματα. Αυτό σημαίνει ότι δεν απαιτείται από το χρήστη να εισάγει κάποια παράμετρο που να ορίζει το επιθυμητό μέγεθος. Η ύπαρξη τέτοιου είδους παραμέτρων καθιστά τη διαδικασία της προ-επεξεργασίας ως μία επίπονη και χρονοβόρα διαδικασία, αφού πρέπει αναζητηθούν και να τους ανατεθούν οι καταλληλότερες τιμές, δηλαδή αυτές που επιτυγχάνουν την υψηλότερη απόδοση. Τέλος, μειονέκτημα του CNN-rule είναι το ότι το συμπυκνωμένο σύνολο δεδομένων που προκύπτει εξαρτάται από την σειρά που αυτά εξετάζονται από τον αλγόριθμο, ή με άλλα λόγια, από τη σειρά που αυτά είναι αποθηκευμένα αρχικό σύνολο εκπαίδευσης. Αυτό σημαίνει ότι είναι πιθανό να παραχθεί διαφορετικό συμπυκνωμένο σύνολο για τα ίδια δεδομένα εκπαίδευσης, αν αυτά αποθηκευτούν με διαφορετική σειρά.

Λαμβάνοντας υπόψη την παραπάνω διαδικασία, γίνεται εύκολα αντιληπτό το πρόβλημα που αποτελεί ο θόρυβος στα δεδομένα εκπαίδευσης. Όσο περισσότερος θόρυβος, τόσα περισσότερα δεδομένα θα μεταφέρονται στο σύνολο Α και συνεπώς τόσο μικρότερος λόγος μείωσης των δεδομένων θα επιτυγχάνεται. Η ανάγκη εφαρμογής μιας διαδικασίας απομάκρυνσης θορύβου, πολλές φορές είναι επιβεβλημένη.

**Αλγόριθμος CNN-rule**

**Είσοδος**: Σύνολο δεδομένων εκπαίδευσης Δ,
πλήθος δεδομένων εκπαίδευσης N

**Έξοδος**: Συμπυκνωμένο σύνολο A

**Αρχή**

1. A[1] ← Δ[1]
2. **Για** i από 2 μέχρι N
3. B[i-1] ← Δ[i]
4. **Τέλος επανάληψης**
5. i ←2
6. **Επανέλαβε**
7. flag ← Ψευδής
8. **Για κάθε** στιγμιότυπο x που ανήκει στο B
9. Αναζήτησε τον εγγύτερο γείτονα k στο x
10. **Αν** η κλάση του x είναι διαφορετική από την κλάση του k **τότε**
11. A[i] ← x
12. i ← i + 1
12. flag ← Αληθής
13. **Τέλος Αν**
14. **Τέλος επανάληψης**
15. **Μέχρις ότου** flag = Ψευδής *!Δεν υπήρξε μετακίνηση από το B στο A*

**Τέλος αλγορίθμου**

Σχήμα 9 "Αλγόριθμος συμπύκνωσης εγγύτερου γείτονα (CNN-rule)"

### 3.1.3 ΑΛΓΟΡΙΘΜΟΣ ΤΩΝ CHEN ΚΑΙ JZWIK

Ένας πολύ γνωστός αλγόριθμος σύνοψης δεδομένων, με πολλές αναφορές στην διεθνή βιβλιογραφία, προτάθηκε το 1996 από τους Chen και Jzwik (αλγόριθμος CJ) (Chen και Jzwik, 1996). Ο αλγόριθμος CJ είναι ο πρόγονος του αλγορίθμου RSP3, ο οποίος αποτελεί αντικείμενο έρευνας στην παρούσα εργασία. Σε αντίθεση με τον αλγόριθμο CNN-rule, ο αλγόριθμος CJ είναι παραμετρικός. Ο χρήστης πρέπει να ορίσει το επιθυμητό μέγεθος του συμπυκνωμένου συνόλου δεδομένων και να το εισάγει στον αλγόριθμο ως παράμετρο.

Ο αλγόριθμος CJ βασίζεται σε μια επαναληπτική διαδικασία διαχωρισμού του αρχικού συνόλου εκπαίδευσης. Συγκεκριμένα, ο αλγόριθμος αρχικά αναζητά τα δυο πιο απομακρυσμένα αντικείμενα στα δεδομένα. Έστω αυτά είναι τα αντικείμενα Α και Β. Στη συνέχεια, το σύνολο δεδομένων χωρίζεται στα δύο υποσύνολα. Τα αντικείμενα που είναι πιο κοντά στο Α αποθηκεύονται στο υποσύνολο $\Sigma_{A,}$ ενώ τα στιγμιότυπα που βρίσκονται πιο κοντά στο Β αποθηκεύονται στο υποσύνολο $\Sigma_B$. Η διαδικασία εύρεσης των πιο απομακρυσμένων αντικειμένων και η διαίρεση των συνόλων συνεχίζεται μέσα στα υποσύνολα $\Sigma_A$ και $\Sigma_B$. Ως κριτήριο για το πιο υποσύνολο θα διαιρεθεί πρώτο, ο αλγόριθμος CJ υιοθετεί το κριτήριο της μεγαλύτερης διαμέτρου. Δηλαδή, το υποσύνολο με την μεγαλύτερη διάμετρο που ορίζεται από τα δύο πιο απομακρυσμένα αντικείμενα διαιρείται πρώτο. Η διαδικασία συνεχίζεται μέχρι ο αριθμός των υποσυνόλων που έχουν δημιουργηθεί είναι ίσος με την παράμετρο που έχει ορίσει ο χρήστης.

Όταν ολοκληρωθεί αυτή η διαδικασία, ο αλγόριθμος βρίσκει την πλειοψηφούσα κλάση σε κάθε υποσύνολο. Στη συνέχεια δημιουργεί μια σύνοψη των αντικειμένων που ανήκουν στην συγκεκριμένη κλάση, βρίσκοντας το μέσο όρο των χαρακτηριστικών (attributes) τους. Τα νέα αντικείμενα που δημιουργούνται με αυτή την διαδικασία απαρτίζουν το συμπυκνωμένο σύνολο δεδομένων. Σημειώνεται ότι τα αντικείμενα που δεν ανήκουν στην πλειοψηφούσα κλάση του κάθε υποσυνόλου, δεν λαμβάνονται υπόψη κατά των υπολογισμό του μέσου όρου και δεν αντιπροσωπεύονται στο συμπυκνωμένο σύνολο. Το σκεπτικό για την υιοθέτηση της διαμέτρου ως κριτήριο επιλογής επόμενου υποσυνόλου, είναι ότι ένα υποσύνολο με μεγαλύτερη διάμετρο είναι πιθανό να περιλαμβάνει περισσότερα αντικείμενα από ότι ένα υποσύνολο με μικρότερη διάμετρο. Με αυτό τον τρόπο, ο αλγόριθμος φιλοδοξεί ότι θα επιτύχει μεγαλύτερη μείωση στον όγκο των δεδομένων.

Τέλος, αξίζει να αναφερθεί ότι το συμπυκνωμένο σύνολο που παράγει ο αλγόριθμος CJ δεν εξαρτάται από τη σειρά που θα εξετάσει τα δεδομένα εκπαίδευσης. Αυτό σημαίνει ότι ο αλγόριθμος καταλήγει στο ίδιο συμπυκνωμένο σύνολο ανεξάρτητα με τη σειρά που είναι αποθηκευμένα τα αντικείμενα στο αρχικό σύνολο εκπαίδευσης.

Η διαδικασία του αλγόριθμου CJ μπορεί να γίνει ευκολότερα κατανοητή μελετώντας το σχήμα 10. Έστω ότι το αρχικό σύνολο δεδομένων διαθέτει 9 αντικείμενα τα οποία κατανέμονται σε δύο κλάσεις, την κλάση «τετράγωνο» και την

κλάση «κύκλος». Επίσης, έστω ότι το επιθυμητό μέγεθος του συμπυκνωμένου συνόλου δεδομένων είναι 6 αντικείμενα. Αρχικά, τα δύο πιο απομακρυσμένα στιγμιότυπα είναι το a και i. Συνεπώς, η πρώτη διαίρεση έχει ως αποτέλεσμα την δημιουργία των συνόλων:

$$\Sigma_1 = \{a, b, c, d, e\} \text{ και } \Sigma_2 = \{f, g, h, i\}.$$

Η ίδια διαδικασία συνεχίζεται και το υποσύνολο $\Sigma_1$ διαιρείτε αφού έχει μεγαλύτερη διάμετρο από το $\Sigma_2$. Στο σημείο αυτό τα υποσύνολα είναι τα:

$$\Sigma_1 = \{a, b, c\}, \Sigma_3 = \{d, e\} \text{ και } \Sigma_2 = \{f, g, h, i\}$$

Το υποσύνολο με τη μεγαλύτερη διάμετρο είναι το $\Sigma_2$. Όποτε διαιρείται. Τα υποσύνολα είναι:

$$\Sigma_1 = \{a, b, c\}, \Sigma_3 = \{d, e\}, \Sigma_2 = \{f, g\} \text{ και } \Sigma_4 = \{h, i\}$$

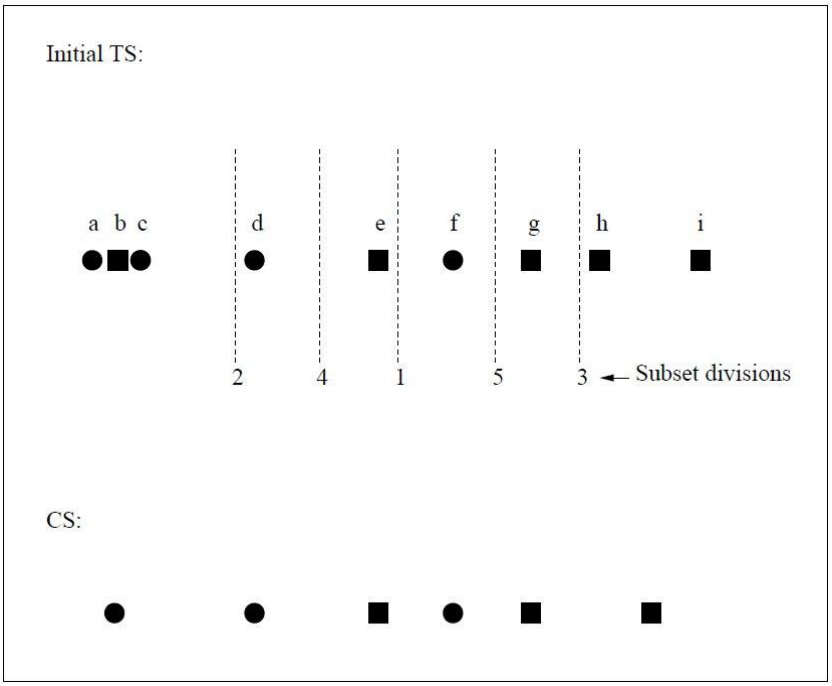
Στο επόμενο βήμα, το υποσύνολο με τη μεγαλύτερη διάμετρο είναι το Σ3. Οπότε:

$$\Sigma_1 = \{a, b, c\}, \Sigma_3 = \{d\}, \Sigma_5 = \{e\} \Sigma_2 = \{f, g\} \text{ και } \Sigma_4 = \{h, i\}$$

Απαιτείται ένα ακόμη υποσύνολο για να καλυφθεί το μέγεθος του επιθυμητού μεγέθους του συμπυκνωμένου συνόλου δεδομένων. Η μεγαλύτερη διάμετρος διαπιστώνεται στο υποσύνολο $\Sigma_2$. Οπότε:

$$\Sigma_1 = \{a, b, c\}, \Sigma_3 = \{d\}, \Sigma_5 = \{e\} \Sigma_2 = \{f\}, \Sigma_6 = \{g\} \text{ και } \Sigma_4 = \{h, i\}$$

Το επόμενο βήμα είναι η δημιουργία του συμπυκνωμένου συνόλου (CS). Το υποσύνολο $\Sigma_1$ δεν είναι ομοιογενές. Οπότε το στιγμιότυπο b αγνοείτε αφού δεν ανήκει στην πλειοψηφούσα κλάση στο $\Sigma_1$. Έτσι, η σύνοψη των a και b αποθηκεύεται στο CS. Αντίστοιχα, η σύνοψη των h και i επίσης αποθηκεύεται στο CS. Τα υπόλοιπα υποσύνολα έχουν μόνο ένα στοιχείο το καθένα. Τα στοιχεία αυτά μεταφέρονται στο CS. Με τον τρόπο αυτό, δημιουργήθηκε ένα συμπυκνωμένο σύνολο δεδομένων 6 αντικειμένων.




Σχήμα 10 "Διαδικασία του αλγόριθμου των Chen και Jzwik"

### 3.1.4 ΟΙΚΟΓΕΝΕΙΑ ΑΛΓΟΡΙΘΜΩΝ ΜΕΙΩΣΗΣ ΔΕΔΟΜΕΝΩΝ ΕΚΠΑΙΔΕΥΣΗΣ ΜΕ ΚΑΤΑΤΜΗΣΗ ΧΩΡΟΥ (RSP ALGORITHMS)

Οι οικογένεια των αλγόριθμων μείωσης δεδομένων εκπαίδευσης με κατάτμηση χώρου (Reduction by Space Partitioning algorithm) (Sanchez, 2004) αποτελείται από τρεις αλγορίθμους, RSP1, RSP2 και RSP3, οι οποίοι αποτελούν επεκτάσεις – παραλλαγές της μεθόδου των Chen και Jzwik. Οι αλγόριθμοι RSP1 και RSP2 είναι παραμετρικοί όπως ο αλγόριθμος CJ. Δηλαδή, δέχονται ως παράμετρο το επιθυμητό μέγεθος του συμπυκνωμένου συνόλου δεδομένων. Αντίθετα, ο αλγόριθμος RSP3 ορίζει αυτόματα το μέγεθος του συμπυκνωμένου συνόλου ανάλογα με το πως κατανέμονται τα δεδομένα στον πολυδιάστατο χώρο καθώς επίσης και με τα επίπεδα θορύβου στα δεδομένα.

Ο αλγόριθμος RSP1 διαφέρει σε σχέση με τον αλγόριθμο CJ σε μια μικρή λεπτομέρεια. Ο RSP1, στο σημείο δημιουργίας του συμπυκνωμένου συνόλου, δημιουργεί τόσες συνόψεις αντικειμένων όσες και οι διαφορετικές κλάσεις σε κάθε υποσύνολο. Στο παράδειγμα που παρουσιάστηκε στην προηγούμενη παράγραφο (Σχήμα 10), ο αλγόριθμος RSP1 θα δημιουργούσε δύο συνόψεις για το υποσύνολο $\Sigma_1$. Ο RSP1 παράγει μεγαλύτερο συμπυκνωμένο σύνολο σε σχέση με τον αλγόριθμο CJ. Ωστόσο, λαμβάνει υπόψη όλα τα αντικείμενα και έτσι ο k-NN κατηγοριοποιητής επιτυγχάνει συνήθως υψηλότερη ακρίβεια.

Το Σχήμα 11 παρουσιάζει τον αλγόριθμο RSP1. Όλα τα βήματα του αλγορίθμου είναι ίδια με αυτά του αλγόριθμου CJ. Εξαίρεση αποτελούν οι γραμμές

19 – 22. Στο σημείο αυτό, ο αλγόριθμος CJ προβλέπει τον εντοπισμό της πλειοψηφούσες κλάσης και τη δημιουργία μιας σύνοψης για αυτή την κλάση.

Ο αλγόριθμος RSP2 διαφέρει σε σχέση με τον RSP1 στον τρόπο επιλογής του επόμενου υποσυνόλου που θα διαιρεθεί. Το σκεπτικό του αλγορίθμου CJ και του RSP1, όπως έχει ήδη ειπωθεί, είναι ότι το υποσύνολο με την μεγαλύτερη διάμετρο, πιθανότατα θα διαθέτει περισσότερα αντικείμενα και έτσι επιτυγχάνεται μεγαλύτερη μείωση όγκου. Ωστόσο αυτό δεν είναι ισχύει πάντα. Υπάρχουν περιπτώσεις όπου υποσύνολα με μικρή διάμετρο περιλαμβάνουν πολλαπλάσια αντικείμενα (πυκνά υποσύνολα) σε σχέση με υποσύνολα με μεγαλύτερη διάμετρο. Ο αλγόριθμος RSP2 εισάγει ένα νέο κριτήριο διαίρεσης. Συγκεκριμένα, επιλέγει το υποσύνολο με την μεγαλύτερη επικάλυψη. Στόχος του κριτηρίου είναι η επίτευξη μεγαλύτερης μείωσης του όγκου των δεδομένων. Όλα τα υπόλοιπα βήματα του RSP2 είναι ίδια με αυτά του RSP1.

Τέλος, ο αλγόριθμος RSP3 δεν υιοθετεί κάποιο κριτήριο για τη διαίρεση των υποσυνόλων ή πιο συγκεκριμένα, δεν είναι σημαντικό το ποίο υποσύνολο θα διασπαστεί πρώτο. Ο RSP3 υιοθετεί την έννοια της ομοιογένειας των υποσυνόλων. Η διαδικασία της διαίρεσης ενός υποσυνόλου συνεχίζεται μέχρις ότου τα υποσύνολα που προκύπτουν είναι ομοιογενή, δηλαδή να περιλαμβάνουν αντικείμενα της ίδια κλάσης. Με αυτό τον τρόπο ο RSP3 καθορίζει αυτόματα το μέγεθος του παραγόμενου συμπυκνωμένου συνόλου δεδομένων και έτσι οι επαναλαμβανόμενες δοκιμές για τον καθορισμού παραμέτρων αποφεύγονται. Φυσικά, το πλήθος των υποσυνόλων που παράγονται, καθώς επίσης και το μέγεθος του συμπυκνωμένου συνόλου, εξαρτάται σε μεγάλο βαθμό από τα επίπεδα θορύβου στα δεδομένα. Επίσης, μελετώντας τον αλγόριθμο γίνεται αντιληπτό ότι αυτός δημιουργεί πολλές συνόψεις αντικειμένων για τις περιοχές δεδομένων που βρίσκονται κοντά σε σύνορα απόφασης. Αντίθετα, δημιουργεί ελάχιστες συνόψεις για τις εσωτερικές περιοχές των κλάσεων, οι οποίες συνήθως είναι ομοιογενείς. Το σχήμα 12 παρουσιάζει τα βήματα του αλγορίθμου. Παρατηρούμε ότι οι διαφορές με τον αλγόριθμο RSP1 του σχήματος 11 είναι πολλές.

Στην βιβλιογραφία αναφέρεται ότι ο αλγόριθμος RSP3 καταλήγει σε μεγαλύτερο συμπυκνωμένο σύνολο σε σχέση με τον CNN-rule. Ωστόσο, ο k-NN κατηγοριοποιητής, που αναζητά για εγγύτερους γείτονες στο συμπυκνωμένο σύνολο του RSP3, συνήθως επιτυγχάνει υψηλότερη ακρίβεια. Τα παραπάνω επιβεβαιώθηκαν από τα πειράματα που εκτελέστηκαν στα πλαίσια αυτής της εργασίας και παρουσιάζονται σε επόμενο κεφάλαιο.

Η αναζήτηση των δυο πιο απομακρυσμένων αντικειμένων κάθε υποσυνόλου αποτελεί μια διαδικασία μεγάλου κόστους, αφού προϋποθέτει τον υπολογισμό όλων των αποστάσεων μεταξύ των αντικειμένων του κάθε υποσυνόλου. Έτσι, οι τρεις αλγόριθμοι RSP και ο CJ συνήθως έχουν υψηλό υπολογιστικό κόστος προ-επεξεργασίας. Σε πολλές εφαρμογές αυτό δεν είναι πρόβλημα αφού η δημιουργία του συμπυκνωμένου συνόλου πραγματοποιείται μια φορά. Από την άλλη, υπάρχουν εφαρμογές και συστήματα που είτε έχουν περιορισμούς χρόνου (time constraint), είτε συχνά λαμβάνουν νέα δεδομένα εκπαίδευσης και η

επαναλαμβανόμενη δημιουργία του συμπυκνωμένου συνόλου δεδομένων είναι επιβεβλημένη. Σε τέτοιου είδους περιβάλλοντα, το υψηλό κόστος επεξεργασίας αποτελεί πρόβλημα και η εκτέλεση των διαδικασιών μείωσης όγκου ίσως είναι κάτι το απαγορευτικό. Το κόστος προ-επεξεργασίας των αλγόριθμων RSP3 και CNN-rule εκτιμάτε και παρουσιάζεται σε επόμενο κεφάλαιο.

Τέλος, σημειώνεται ότι το συμπυκνωμένο σύνολο δεδομένων που προκύπτει από τους αλγόριθμους RSP και CJ, σε αντίθεση με αυτό του CNN-rule, δεν εξαρτάται από την σειρά που εξετάζονται τα δεδομένα εκπαίδευσης.

**Αλγόριθμος RSP1**

**Είσοδος**: Σύνολο δεδομένων εκπαίδευσης Δ,

Επιθυμητό μέγεθος συμπυκνωμένου συνόλου Μ

**Έξοδος**: Συμπυκνωμένο σύνολο CS

**Αρχή**

1. L ← Δ *! L: Λίστα υποσυνόλων*
2. πλήθος ← 1 *! πλήθος υποσυνόλων*
3. **Όσο** πλήθος ≤ Μ **επανέλαβε**
4. Τ ← Υποσύνολο της λίστας L με την μεγαλύτερο διάμετρο
5. Υπολόγισε τις αποστάσεις μεταξύ των στιγμιότυπων στο Τ
6. Βρες την μεγαλύτερη απόσταση (διάμετρο) στο Τ και τα στιγμιότυπα Α και Β που την ορίζουν
7. **Για κάθε** στιγμιότυπο x στο Τ
8. **Αν** απόσταση (x, A) ≤ απόσταση (x, B) **τότε**
9. $\Sigma_A$ ← $\Sigma_A$ U x
10. **Αλλιώς**
11. $\Sigma_B$ ← $\Sigma_B$ U x
12. **Τέλος Αν**
13. **Τέλος επανάληψης**
14. L ← L - T
15. L ← L U $\Sigma_A$ U $\Sigma_B$
16. πλήθος ← πλήθος + 1
17. **Τέλος επανάληψης**
18. Για κάθε υποσύνολο Τ στη λίστα L
19. Για κάθε διαφορετική κλάση Κ στο Τ
20. Σ ← Δημιουργία σύνοψης: Υπολόγισε το μέσο όρο των γνωρισμάτων των στιγμιότυπων του Τ που ανήκουν στην κλάση Κ
21. CS ← CS U Σ
22. Τέλος επανάληψης
23. Τέλος επανάληψης
24. επέστρεψε CS

**Τέλος Αλγορίθμου**

Σχήμα 11 "Αλγόριθμος RSP1"

**Αλγόριθμος RSP3**

**Είσοδος**: Σύνολο δεδομένων εκπαίδευσης Δ

**Έξοδος**: Συμπυκνωμένο σύνολο CS

**Αρχή**

1. L ← Δ *! L: Λίστα υποσυνόλων*
2. **Επανέλαβε**
3. Τ ← Υποσύνολο της λίστας L με την μεγαλύτερο διάμετρο
4. **Αν** Τ είναι ομοιογενές **τότε**
5. Σ ← Δημιουργία σύνοψης: Υπολόγισε το μέσο όρο των γνωρισμάτων των στιγμιότυπων του Τ
6. CS ← CS U Σ
7. **Αλλιώς**
8. Υπολόγισε τις αποστάσεις μεταξύ των στιγμιότυπων στο Τ
9. Βρες την μεγαλύτερη απόσταση (διάμετρο) στο Τ και τα στιγμιότυπα Α και Β που την ορίζουν
10. **Για κάθε** στιγμιότυπο x στο Τ
11. **Αν** απόσταση (x, A) ≤ απόσταση (x, B) **τότε**
12. $\Sigma_A$ ← $\Sigma_A$ U x
13. **Αλλιώς**
14. $\Sigma_B$ ← $\Sigma_B$ U x
15. **Τέλος Αν**
16. **Τέλος επανάληψης**
17. **Τέλος Αν**
18. L ← L - T
19. L ← L U $\Sigma_A$ U $\Sigma_B$
20. **Μέχρι ότου** κάθε υποσύνολο της L είναι ομοιογενείς.
21. επέστρεψε CS

**Τέλος Αλγορίθμου**

Σχήμα 12 "Αλγόριθμος RSP3"

### 3.1.5 Ο ΚΑΝΟΝΑΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΓΓΥΤΕΡΟΥ ΓΕΙΤΟΝΑ (ENN-rule)

Όπως έχει αναφερθεί αναλυτικά στην παράγραφο 3.1.1, το ποσοστό μείωσης του όγκου των δεδομένων που επιτυγχάνουν οι αλγόριθμοι που μελετάμε, εξαρτάται σε μεγάλο βαθμό από τα επίπεδα θορύβου στα δεδομένα εκπαίδευσης. Έτσι, πολλές φορές είναι απαραίτητο να εκτελεστεί ένας αλγόριθμος επεξεργασίας (editing). Αυτός απομακρύνει τα δεδομένα που αποτελούν θόρυβο πριν την εκτέλεση των αλγορίθμων μείωσης όγκου όπως ο CNN-rule και ο RSP3. Οι αλγόριθμοι επεξεργασίας, εκτός από την απομάκρυνση του θορύβου, στοχεύουν στην εξομάλυνση των συνόρων απόφασης ώστε αυτά να ορίζουν χωρίς αμφιβολία τα όρια των περιοχών των κλάσεων.

Ο πρώτος και ένας από τους πιο γνωστούς αλγορίθμους επεξεργασίας είναι ο αλγόριθμος, που προτάθηκε από τον Wilson το 1972 (Wilson, 1972). Ο αλγόριθμος του Wilson είναι γνωστός στην βιβλιογραφία με το όνομα αλγόριθμος επεξεργασίας εγγύτερου γείτονα (Edited Nearest Neighbor Rule – ENN-rule). Ο ENN-rule βασίζεται σε μια απλή ιδέα: ένα αντικείμενο αποτελεί θόρυβο αν οι εγγύτεροι γείτονες του ανήκουν σε άλλη κλάση. Φυσικά υπάρχουν πολλοί άλλοι αλγόριθμοι επεξεργασίας. Πολλοί από αυτούς αποτελούν επεκτάσεις ή παραλλαγές του ENN-rule. Για παράδειγμα ο αλγόριθμος all-k-NN (Tomek, 1976) εφαρμόζει επαναληπτικά τον αλγόριθμο του Wilson. Ωστόσο, ο ENN-rule εξακολουθεί να είναι ένας από τους πιο αποδοτικούς αλγορίθμους επεξεργασίας και εφαρμόζεται σε πειράματα από πολλές εργασίες στη βιβλιογραφία. Επίσης, πρέπει να αναφερθεί ότι έχουν προταθεί αλγόριθμοι που υιοθετούν την απομάκρυνση θορύβου στη βασική διαδικασία μείωσης όγκου των δεδομένων εκπαίδευσης. Παράδειγμα τέτοιων αλγορίθμων είναι οι αλγόριθμοι: PGF (Lam et al., 2002), IB3 (Aha et al., 1991) και DROP3 (Wilson και Martinez, 2000).

Ο αλγόριθμος ENN-rule, για κάθε αντικείμενο x του αρχικού συνόλου εκπαίδευσης, αναζητά τους k εγγύτερους γείτονες στο ίδιο σύνολο. Αν η πλειοψηφούσα κλάση των k εγγύτερων γειτόνων είναι διαφορετική από την κλάση του στιγμιότυπου x, το x διαγράφεται αφού εκτιμάται από τον αλγόριθμο ότι αναπαριστά θόρυβο. Στο τέλος το επεξεργασμένο σύνολο δεδομένων περιλαμβάνει τα αντικείμενα που δεν διαγράφηκαν (δηλ. αυτά που η κλάση τους είναι ίδια με την πλειοψηφούσα κλάση). Η προ-περιγραφείσα διαδικασία του αλγόριθμου ENN-rule συνοψίζεται στο Σχήμα 13.

Όπως γίνεται εύκολα αντιληπτό, ο ENN-rule υπολογίζει όλες τις πιθανές αποστάσεις μεταξύ των δεδομένων εκπαίδευσης. Συγκεκριμένα, για ένα σύνολο N αντικειμένων, ο αλγόριθμος υπολογίζει $\frac{N \cdot (N-1)}{2}$ αποστάσεις. Αν και ο αριθμός αυτός αντικατοπτρίζει ένα αρκετά μεγάλο κόστος επεξεργασίας, το οποίο προστίθεται στο κόστος προ-επεξεργασίας που εισάγουν οι αλγόριθμοι μείωσης όγκου, η απομάκρυνση του θορύβου από τα δεδομένα είναι μια απαραίτητη διαδικασία, όχι μόνο για την αποδοτική εφαρμογή των αλγορίθμων μείωσης όγκου,

αλλά γενικότερα για την αποδοτικότερη εφαρμογή διαφόρων αλγορίθμων και τεχνικών εξόρυξης γνώσης.

Τέλος, ένα από τα θέματα που χρίζει αντιμετώπισης για την εφαρμογή του αλγορίθμου του Wilson είναι ο καθορισμός της τιμής της παραμέτρου k. Δηλαδή πόσοι εγγύτεροι γείτονες θα καθορίσουν την πλειοψηφούσα κλάση. Στην βιβλιογραφία γίνεται λόγος ότι η παράμετρος αυτή πρέπει να είναι ένας μικρός και περιττός αριθμός. Φυσικά για την απόφαση στην επιλογή του k παίζει ρόλο το πλήθος των διακριτών κλάσεων στα δεδομένα. Οι Wilson και Martinez (Wilson και Martinez, 2000) προτείνουν ως καταλληλότερη τιμή για την παράμετρο την τιμή k=3. Η τιμή αυτή θα υιοθετηθεί και στα πειράματα της παρούσας εργασίας.

**Αλγόριθμος ENN-rule**

**Είσοδος**: Σύνολο δεδομένων εκπαίδευσης Δ

Παράμετρος k

**Έξοδος**: Επεξεργασμένο σύνολο ES

**Αρχή**

1. ES ← Δ
2. **Για κάθε** στιγμιότυπο x με κλάση T στο Δ
3. Αναζήτησε τους k εγγύτερους γείτονες του x στο Δ βάσει ενός μέτρου απόστασης
4. M ← Πλειοψηφούσα κλάση των k εγγύτερων γειτόνων
5. **Αν** T ≠ M **τότε**
6. ES ← ES - x
7. **Τέλος Αν**
8. **Τέλος επανάληψης**
9. Επέστρεψε ES

**Τέλος Αλγορίθμου**

Σχήμα 13 "Ο αλγόριθμος του Wilson (ENN-rule) "

## 3.2 ΤΕΧΝΙΚΕΣ ΜΕΙΩΣΗΣ ΠΛΗΘΟΥΣ ΔΙΑΣΤΑΣΕΩΝ (ΤΜΠΔ)

### 3.2.1 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Το κόστος επεξεργασίας του αλγόριθμου κατηγοριοποίησης k-NN, εκτός από το πλήθος των δεδομένων εκπαίδευσης, εξαρτάται και από το πλήθος των διαστάσεων (χαρακτηριστικών) των δεδομένων και αυτό γιατί ο υπολογισμός της απόστασης δύο αντικειμένων περιλαμβάνει περισσότερους υπολογισμούς αφού είναι ανάλογος αυτών. Το υψηλό κόστος υπολογισμού αποστάσεων εντοπίζεται κυρίως σε δεδομένα χρονοσειρών, στα οποία επικεντρώνεται και η παρούσα εργασία. Μία χρονοσειρά μήκους n αποτελεί ένα αντικείμενο αντίστοιχου πλήθους διαστάσεων. Είναι φανερό ότι όσο το n παίρνει μεγαλύτερες τιμές, τόσο η εφαρμογή του κατηγοριοποιητή τείνει να υλοποιείται λιγότερο αποδοτικά και αποτελεσματικά. Συνεπώς είναι αναγκαία η αναπαράσταση της χρονοσειράς σε μια μορφή, η οποία θα ορίζεται από λιγότερες διαστάσεις και θα συμπιέζει τα αρχικά δεδομένα σε μικρότερο όγκο.

Αν και έχουν προταθεί διάφορες ΤΜΠΔ, όπως η ανάλυση κυρίων συνιστωσών (Principal Component Analysis - PCA) (Jolliffe, 2002), στην εργασία αυτή επικεντρωνόμαστε σε αυτές που έχουν προταθεί για την αναπαράσταση χρονοσειρών σε χώρο λιγότερων διαστάσεων. Για να είναι αποτελεσματική μια τέτοια αναπαράσταση θα πρέπει να εγγυάται ότι: (i) θα διατηρεί τα σημαντικότερα χαρακτηριστικά-πληροφορίες της αρχικής χρονοσειράς και (ii) θα συμπιέζει τα αρχικά δεδομένα σε ικανοποιητικό βαθμό (Agrawal et al, 1992).

Στα πλαίσια της ΤΜΠΔ σε χρονοσειρές, οι αναπαραστάσεις οι οποίες έχουν προταθεί και μελετηθεί περισσότερο είναι οι εξής:

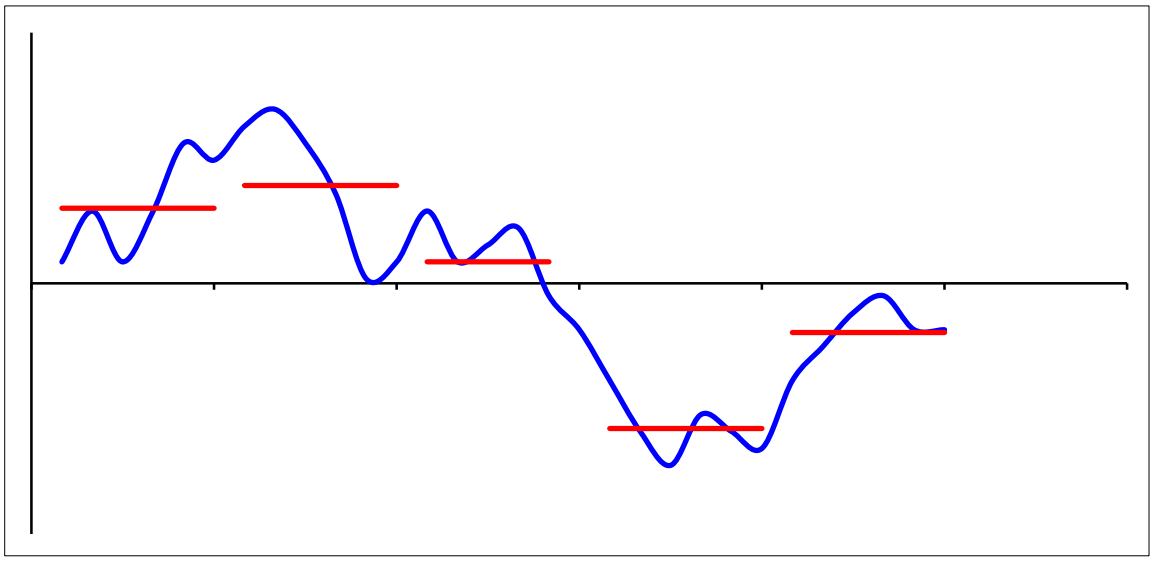
- Τμηματικά Πολυώνυμα (Piecewise Polynomials)
  - Τμηματική Γραμμική Προσέγγιση (Piecewise Linear Approximation)
  - Προσαρμοστική Τμηματική Σταθερή Προσέγγιση (Adaptive Piecewise Constant Approximation)
- Διάσπαση Ιδιαζουσών Τιμών (Singular Value Decomposition)
- Συμβολική (Symbolic)
- Διακριτός Μετασχηματισμός Wavelet
- Διακριτός Μετασχηματισμός Fourier (Discrete Fourier Transform)
- Τμηματική Συνολική Προσέγγιση (Piecewise Aggregate Approximation)

Στην πειραματική μελέτη, που εκπονήθηκε στα πλαίσια της εργασίας, χρησιμοποιήθηκε η Τμηματική Συνολική Προσέγγιση για την πιο αποδοτική εκτέλεση του κατηγοριοποιητή k εγγύτερων γειτόνων .

### 3.2.1 ΤΜΗΜΑΤΙΚΗ ΣΥΝΟΛΙΚΗ ΠΡΟΣΕΓΓΙΣΗ (PIECEWISE AGGREGATE APPROXIMATION)

Η μέθοδος της Τμηματικής Συνολικής Προσέγγισης (Keogh and Pazzani, 2000), (Keogh, 2001), (Yi and Faloutsos, 2000) διαχωρίζει μία χρονοσειρά σε τμήματα ίσου μήκους και υπολογίζει την αντίστοιχη μέση τιμή κάθε ενός από αυτά. Η ακολουθία των μέσων αυτών τιμών αποτελεί την αναπαράσταση της αρχικής χρονοσειράς.

Παρόλη την απλότητά της, η προσέγγιση αυτή έχει αποδειχθεί εξίσου αποδοτική με τις επικρατέστερες μεθόδους αναπαράστασης, οι απαιτούμενοι υπολογισμοί επιτυγχάνονται τάχιστα και ταυτόχρονα υποστηρίζει αποτελεσματικά τα σημαντικότερα μέτρα ομοιότητας.



Σχήμα 14 "Αναπαράσταση μείωσης διαστάσεων τμηματικής συνολικής προσέγγισης"

## ΕΠΙΛΟΓΟΣ

Η μείωση όγκου δεδομένων για την ταχύτερη κατηγοριοποίηση με χρήση του αλγόριθμου k-NN είναι ένα πολύ σημαντικό πεδίο έρευνας. Αυτό αποδεικνύεται από το ότι πολλοί ερευνητές έχουν στρέψει το ενδιαφέρον τους σε αυτό το πεδίο, με αποτέλεσμα την δημοσίευση πολλών εργασιών που προτείνουν αντίστοιχες μεθόδους. Στο κεφάλαιο αυτό έγινε μια σύντομη αναφορά σε τέτοιου είδους μεθόδους, καθώς επίσης παρουσιάστηκαν αναλυτικά κάποιοι από τους πιο γνωστούς αλγορίθμους. Η απόδοση τριών αλγορίθμων που παρουσιάστηκαν πρόκειται να μας απασχολήσει στο επόμενο κεφάλαιο.

# ΚΕΦΑΛΑΙΟ 4

# ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

## ΕΙΣΑΓΩΓΗ

Οι τεχνικές μείωσης δεδομένων εκπαίδευσης (ΤΜΔΕ) που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, αν και έχουν εφαρμοστεί με επιτυχία σε μεγάλα σύνολα δεδομένων, δεν έχουν εφαρμοστεί σε δεδομένα χρονοσειρών. Φυσικά στη βιβλιογραφία συναντάμε ΤΜΔΕ που έχουν προταθεί αποκλειστικά για δεδομένα χρονοσειρών (Xi et al., 2006), (Buza et al., 2011). Οι τεχνικές αυτές ανήκουν στην κατηγορία επιλογής δεδομένων. Το βασικό τους μειονέκτημα είναι ότι είναι παραμετρικές. Ο χρήστης πρέπει να ορίσει το πλήθος των χρονοσειρών που θα επιλεγούν. Γνωστές, μη παραμετρικές ΤΜΔΕ επιλογής, όπως ο CNN-rule δεν έχουν εφαρμοστεί σε δεδομένα χρονοσειρών. Επιπρόσθετα, καμιά από τις γνωστές ΤΜΔΕ σύνοψης δεν έχει εφαρμοστεί σε χρονοσειρές. Το κεφάλαιο αυτό περιλαμβάνει μια εκτεταμένη πειρατική μελέτη, όπου μη παραμετρικές ΤΜΔΕ καθώς και τεχνικές μείωσης πλήθους διαστάσεων (ΤΜΠΔ), οι οποίες παρουσιάστηκαν κεφάλαιο 3, εφαρμόζονται σε δεδομένα χρονοσειρών.

## 4.1 ΠΕΡΙΒΑΛΛΟΝ ΠΕΙΡΑΜΑΤΙΚΗΣ ΜΕΛΕΤΗΣ

### 4.1.1 ΠΕΡΙΓΡΑΦΗ

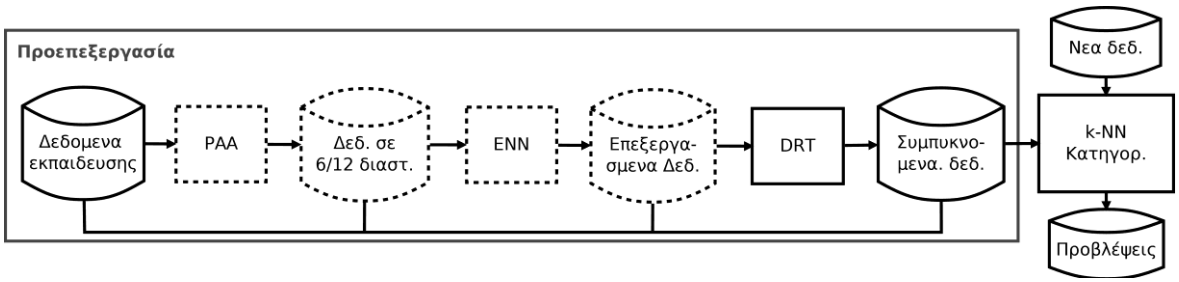
Η πειρατική μελέτη που εκπονήθηκε στα πλαίσια της εργασίας, αποσκοπεί στην μελέτη της απόδοσης δύο δημοφιλών, μη παραμετρικών ΤΜΔΕ σε δεδομένα χρονοσειρών. Συγκεκριμένα οι τεχνικές εκτελέστηκαν σε επτά σύνολα δεδομένων χρονοσειρών και η απόδοση τους συγκρίθηκε εκτιμώντας τέσσερις μετρήσεις εκτίμησης της απόδοσης. Η πειραματική μελέτη επικεντρώνεται σε μια ΤΜΔΕ επιλογής και μια σύνοψης: (i) Κανόνας συμπύκνωσης εγγύτερου γείτονα (CNN-rule) (Hart, 1968) και (ii) Τρίτος αλγόριθμος κατάτμησης χώρου RSP3 (Sanchez, 2004). Αυτές οι τεχνικές μπορούν να θεωρηθούν ως «καλοί» αντιπρόσωποι των δύο κατηγοριών ΤΜΔΕ. Ως μέτρο απόστασης χρησιμοποιήθηκε η ευκλείδεια απόσταση, η οποία για δύο αντικείμενα $m$ διαστάσεων $x_i$ και $x'$, υπολογίζεται αθροίζοντας τις διαφορές των αντιστοιχών διαστάσεων ως εξής:

$$d(x_i, x') = \sqrt{\sum_{j=1}^{m} \left(a_j(x_i) - a_j(x')\right)^2}$$

Όλα τα πειράματα εκτελέστηκαν δύο φορές: μία στα πραγματικά δεδομένα και μία σε δεδομένα χωρίς θόρυβο, που προέκυψαν μετά από την επεξεργασία τους από την τεχνική του κανόνα επεξεργασίας εγγύτερου γείτονα (ENN-rule) (Wilson, 1972). Τέλος, πειράματα εκτελέστηκαν στα συγκεκριμένα σύνολα δεδομένα χρονοσειρών, αφού πρώτα μετασχηματίστηκαν σε χώρο με λιγότερες διαστάσεις. Συγκεκριμένα, όλα τα σύνολα δεδομένων μετασχηματίστηκαν σε χώρο 6 και 12 διαστάσεων. Για τον μετασχηματισμό, χρησιμοποιήθηκε η τεχνική τμηματικής συνολικής προσέγγισης (Piecewise Aggregate Approximation - PAA) (Keogh and Pazzani, 2000), (Keogh, 2001), (Yi and Faloutsos, 2000). Η πειραματική διαδικασία που ακολουθήθηκε είναι η ίδια με αυτή των αρχικών συνόλων δεδομένων.

Σε όλες τις περιπτώσεις, οι αποδόσεις των ΤΜΔΕ συγκρίθηκαν με αυτή του «συμβατικού» κατηγοριοποιητή εγγύτερων γειτόνων, δηλαδή τον κατηγοριοποιητή που εξετάζει το αρχικό σύνολο δεδομένων (χωρίς καμία προεπεξεργασία) για να προβλέψει την κλάση νέων αντικειμένων.

Σημειώνεται ότι οι ΤΜΔΕ CNN-rule, RSP3 και ENN-rule, ο αλγόριθμος k-NN καθώς και η ΤΜΠΔ PAA έχουν υλοποιηθεί από μέλη της ομάδας διαχείρισης πληροφορίας[1] του τμήματος πανεπιστημίου Μακεδονίας, μέλος της οποίας είναι και ο επιβλέπων της εργασίας. Στα πλαίσια της εκπόνησης της πειρατικής μελέτης αυτής της εργασίας, χρησιμοποιήθηκαν οι εν λόγω υλοποιήσεις. Φυσικά, οι πηγαίοι κώδικες των ΤΜΔΕ τροποποιήθηκαν σε πολλά σημεία ώστε αυτές να μπορούν να χειριστούν σύνολα δεδομένων χρονοσειρών.




Σχήμα 15 "Σχηματική αναπαράσταση πειραματικής διαδικασίας"

### 4.1.2 ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΧΡΟΝΟΣΕΙΡΩΝ

Όλα τα πειράματα εκτελέστηκαν σε δεδομένα χρονοσειρών που ανακτήθηκαν από το τον διαδικτυακό τόπο UCR[2]. Ο διαδικτυακός τόπος περιλαμβάνει ένα μεγάλο αριθμό από σύνολα δεδομένων χρονοσειρών τα οποία έχουν χρησιμοποιηθεί κατά καιρούς από διάφορους ερευνητές που δραστηριοποιούνται στο χώρο της εξόρυξης γνώσης από χρονοσειρές. Τα

[1] Information Management Lab (IML), University of Macedonia: http://iml.it.uom.gr/

[2] UCR Time Series Classification/Clustering Page: http://www.cs.ucr.edu/~eamonn/time_series_data/

χαρακτηριστικά των επτά συνόλων δεδομένων που χρησιμοποιήθηκαν στην τρέχουσα μελέτη περιγράφονται στον πίνακα 2.

Πίνακας 2 "Σύνολα δεδομένων χρονοσειρών"

| Σύνολο Δεδομένων | Μέγεθος | Διαστάσεις | Κλάσεις |
|---|---|---|---|
| Synthetic Control | 600 | 60 | 6 |
| Face All | 2250 | 131 | 14 |
| Two Patterns | 5000 | 128 | 4 |
| Yoga | 3300 | 426 | 2 |
| Wafer | 7164 | 152 | 2 |
| SwedishLeaf | 1125 | 128 | 15 |
| CBF | 930 | 128 | 3 |

### 4.1.3 ΔΙΑΜΟΡΦΩΣΗ ΣΥΝΟΛΩΝ ΔΕΔΟΜΕΝΩΝ ΧΡΟΝΟΣΕΙΡΩΝ

Η εκτίμηση της απόδοσης των αλγορίθμων που εκτελέστηκαν πραγματοποιήθηκε χρησιμοποιώντας την μέθοδο 5-folds cross validation (βλέπε παράγραφο 1.3). Τα επτά σύνολα χρονοσειρών ανακτήθηκαν σε μορφή που δεν ήταν κατάλληλη για τέτοιο είδος validation. Τα δεδομένα ανακτήθηκαν σε μορφή: δεδομένα εκπαίδευσης (ΔΕ) και δεδομένα δοκιμής (ΔΔ). Αντίθετα, το 5-fold cross validation προϋποθέτει την ύπαρξη πέντε ζευγαριών συνόλων εκπαίδευσης και δοκιμής με αναλογία 80% των δεδομένων στο σύνολο εκπαίδευσης και 20% των δεδομένων στο σύνολο δοκιμής. Για να επιτευχθεί κάτι τέτοιο, για κάθε σύνολο δεδομένων μεγέθους *N*, ακολουθήθηκε η παρακάτω διαδικασία:

- Αρχικά, το ΔΕ και ΔΔ συγχωνεύτηκαν σε ένα ενιαίο σύνολο δεδομένων, έστω ΣΔ, το οποίο περιέχει *N* χρονοσειρές (μέγεθος συνόλων – βλέπε πίνακα 2)
- Στην συνέχεια, οι χρονοσειρές του ΣΔ αποθηκεύτηκαν με τυχαίο τρόπο σε ένα νέο σύνολο δεδομένων, έστω ΣΔΤ. Το ΣΔΤ περιλαμβάνει *N* χρονοσειρές σε τυχαία σειρά και όχι ταξινομημένες με βάση το χαρακτηριστικό κλάσης
- Τέλος, δημιουργήθηκαν πέντε ζευγάρια συνόλων. Το κάθε ζευγάρι έχει ένα σύνολο δεδομένων εκπαίδευσης και ένα δοκιμής. Στο πρώτο ζευγάρι, οι πρώτες *M* χρονοσειρές του ΣΔΤ αποθηκεύτηκαν στο σύνολο δοκιμής $ΔΔ_1$, ενώ οι υπόλοιπες *X* χρονοσειρές αποθηκεύτηκαν στο σύνολο εκπαίδευσης $ΔΕ_1$. Οι αριθμοί *M* και *X* είναι το 20% και 80% των *N* χρονοσειρών. Στο επόμενο ζευγάρι, οι επόμενες M χρονοσειρές του ΣΔΤ (δηλ. από τη χρονοσειρά *M*+1 μέχρι τη *M**2) αποθηκεύτηκαν στο $ΔΔ_2$, ενώ όλα τα

υπόλοιπα (δηλ. από τη χρονοσειρά 1 μέχρι τη χρονοσειρά *M* και από τη *M**2+1 μέχρι την χρονοσειρά *N*) αποθηκεύτηκαν στο σύνολο ΔΕ2. Με αντίστοιχο τρόπο δημιουργήθηκαν και τα άλλα 3 ζευγάρια: $\Delta\Delta_3$ & $\Delta E_3$, $\Delta\Delta_4$ & $\Delta E_4$, $\Delta\Delta_5$ & $\Delta E_5$.

Ο βασικός λόγος εκτέλεσης του 2ου βήματος της παραπάνω διαδικασίας, ήταν ότι κάποια σύνολα δεδομένων διατίθενται από το UCR ταξινομημένα με βάση το χαρακτηριστικό κλάσης. Ωστόσο, το συμπυκνωμένο σύνολο που παράγει ο αλγόριθμος CNN-rule, όπως έχει αναφερθεί στο κεφάλαιο 3, εξαρτάται από την σειρά των χρονοσειρών στο σύνολο. Μια τέτοια ταξινόμηση θα επηρέαζε την απόδοση του και έτσι θεωρήθηκε σκόπιμο οι χρονοσειρές κάθε συνόλου να μπουν σε μια τυχαία σειρά. Όλοι οι υπόλοιποι αλγόριθμοι που χρησιμοποιήθηκαν δεν επηρεάζονται από τη σειρά των χρονοσειρών.

Τέλος, πρέπει να αναφερθεί ότι με βάση την μέθοδο επικύρωσης 5-folds cross validation, οι ΤΜΔΕ εκτελέστηκαν σε κάθε ένα από τα σύνολα εκπαίδευσης $\Delta E_i$, $i$=1,2,…,5. Όλες οι μετρήσεις της απόδοσης που σχετίζονται με τις αντίστοιχες ΤΜΔΕ αποτελούν το μέσο όρο που προκύπτει από τις πέντε ανεξάρτητες μετρήσεις.

### 4.1.4 ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ

Οι αλγόριθμοι CNN-rule και RSP3 είναι μη παραμετρικοί. Κατά την προεπεξεργασία των αρχικών δεδομένων, δηλαδή για την κατασκευή του συμπυκνωμένου συνόλου, δεν απαιτείται να ορίσουμε κάποια παράμετρο. Ωστόσο, στο στάδιο της κατηγοριοποίησης νέων χρονοσειρών, κάθε συμπυκνωμένο (ή μη) σύνολο, προσπελαύνεται από τον κατηγοριοποιητή εγγύτερων γειτόνων. Ο κατηγοριοποιητής αυτός περιλαμβάνει την παράμετρο *k*. Στη βιβλιογραφία έχει αποδειχθεί ότι καταλληλότερη τιμή για δεδομένα χρονοσειρών είναι *k*=1. Αυτό επιβεβαιώθηκε και κατά τη διάρκεια των πειραμάτων μας. Συγκεκριμένα, εκτελέστηκαν επαναλαμβανόμενα πειράματα για διαφορετικές τιμές της παραμέτρου και διαπιστώθηκε ότι για *k*=1 ο κατηγοριοποιητής επιτυγχάνει την υψηλότερη ακρίβεια. Έτσι, κατά τη διάρκεια εκτέλεσης όλων των πειραμάτων, υιοθετήθηκε ο κατηγοριοποιητής 1-NN ως η καλύτερη δυνατή λύση.

Κατά τη διάρκεια εκπόνησης της πειραματικής μελέτης, όλα τα πειράματα εκτελέστηκαν και σε δεδομένα χωρίς θόρυβο, τα οποία προέκυψαν από την εκτέλεση του αλγόριθμου επεξεργασίας ENN-rule. Ο αλγόριθμος αυτός ελέγχει την κλάση των *k* εγγύτερων γειτόνων κάθε αντικειμένου *x* του συνόλου εκπαίδευσης. Αν η κλάση του *x* δεν συμφωνεί με την πλειοψηφία, τότε απομακρύνεται (για λεπτομέρειες, βλέπε παράγραφο 3.1.5). Η τιμή του *k* και σε αυτή την περίπτωση, πρέπει να προσδιοριστεί από το χρήστη. Στην εργασία των Wilson και Martinez (Wilson and Martinez, 2000) αποδεικνύεται ότι κατάλληλη τιμή είναι *k*=3. Βάσει αυτού, υιοθετήθηκε η τιμή *k*=3 για όλα τα πειράματα που εκτελέστηκαν.

Τέλος, η τεχνική τμηματικής συνολικής προσέγγισης (PAA), για την αναπαράσταση χρονοσειρών σε λιγότερες διαστάσεις, προϋποθέτει τον προσδιορισμό του πλήθους των διαστάσεων. Στα πλαίσια αυτής της εργασίας, η μέθοδος PAA εκτελέστηκε δύο φορές για κάθε σύνολο δεδομένων. Μια φορά για την παραγωγή ενός συνόλου με χρονοσειρές που αναπαριστά το αρχικό σύνολο με έξι διαστάσεις και μια με δώδεκα. Οπότε, όλα τα πειράματα πραγματοποιήθηκαν τρεις φορές για κάθε σύνολο που έχουμε στην διάθεση μας: πραγματικές διαστάσεις του συνόλου, έξι και δώδεκα διαστάσεις.

### 4.1.5 ΜΕΤΡΗΣΕΙΣ ΓΙΑ ΤΗΝ ΑΠΟΤΙΜΗΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ

Για κάθε τεχνική που εκτελέστηκε, εκτιμήθηκαν τέσσερις μετρήσεις που αφορούν την απόδοση τους. Δύο μετρήσεις αφορούν το στάδιο της προεπεξεργασίας και δύο της κατηγοριοποίησης των χρονοσειρών.

Κατά την προεπεξεργασία, η οποία στοχεύει στην κατασκευή του συμπυκνωμένου συνόλου χρονοσειρών, εκτιμήθηκε το κόστος προεπεξεργασίας καθώς και το ποσοστό συμπύκνωσης των δεδομένων (όσον αφορά το πλήθος των χρονοσειρών) που επιτεύχθηκε. Το κόστος προεπεξεργασίας υπολογίστηκε μετρώντας πόσες διαφορές χαρακτηριστικών αθροίστηκαν κατά των υπολογισμό των ευκλείδειων αποστάσεων. Στην πράξη, η τιμή αυτή υπολογίστηκε μετρώντας το πλήθος των αποστάσεων που υπολογίστηκαν κατά την προεπεξεργασία πολλαπλασιασμένο με το πλήθος των διαστάσεων. Με τον τρόπο αυτό λαμβάνεται υπόψη και το κόστος που εισάγεται από την αύξηση των διαστάσεων. Τέλος, σημειώνεται ότι το κόστος μείωσης των διαστάσεων με την τεχνική PAA δε λήφθηκε υπόψη. Το κόστος αυτό είναι πολύ μικρό λόγω της απλότητας της τεχνικής.

Στο στάδιο της κατηγοριοποίησης εκτιμήθηκε η ακρίβεια κατηγοριοποίησης, που επιτυγχάνει ο κατηγοριοποιητής 1-NN χρησιμοποιώντας το εκάστοτε συμπυκνωμένο ή μη σύνολο, καθώς και το κόστος κατηγοριοποίησης. Το κόστος αυτό, αν και είναι αλληλένδετο με το μέγεθος του συμπυκνωμένου συνόλου που κατασκευάστηκε κατά την προεπεξεργασία, εξαρτάται και από το πλήθος των διαστάσεων. Το κόστος κατηγοριοποίησης υπολογιστικέ με αντίστοιχο τρόπο που υπολογίστηκε και το κόστος προεπεξεργασίας. Υπενθυμίζουμε ότι αν έχουμε *n* χρονοσειρές που πρέπει να κατηγοριοποιηθούν και το σύνολο εκπαίδευσης περιλαμβάνει *m* χρονοσειρές, ο 1-NN κατηγοριοποιητής υπολογίζει *m* x *n* αποστάσεις.

Συνοψίζοντας, οι μετρήσεις που εκτιμήθηκαν είναι κατά την διάρκεια διεξαγωγής της πειρατικής μελέτης είναι:

- Κόστος προεπεξεργασίας (preprocessing cost)
- Λόγος μείωσης (reduction rate)
- Ακρίβεια κατηγοριοποίησης (classification accuracy)
- Κόστος κατηγοριοποίησης (classification cost)

## 4.2 ΑΠΟΤΙΜΗΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ - ΣΥΓΚΡΙΣΕΙΣ

### 4.2.1 ΣΥΝΟΛΙΚΑ

Οι πίνακες 3-8 παρουσιάζουν τις μετρήσεις που προέκυψαν από την εκτέλεση των πειραμάτων. Υπενθυμίζουμε ότι όλες οι μετρήσεις αφορούν μέσους όρους που προέκυψαν από την διαδικασία 5-folds cross validation. Κάθε πίνακας περιλαμβάνει τρεις γραμμές για κάθε σύνολο δεδομένων χρονοσειρών. Κάθε μια από τις τρεις γραμμές αφορά διαφορετικό πλήθος διαστάσεων.

Είναι φανερό ότι το υπολογιστικό κόστος είτε είναι κατηγοριοποίησης, είτε προεπεξεργασίας εξαρτάται σε μεγάλο βαθμό από το πλήθος των διαστάσεων (attributes). Βασικός σκοπός των πειραμάτων είναι να εξακριβωθεί το κατά πόσο η ακρίβεια κατηγοριοποίησης των μεθόδων επηρεάζεται από την μείωση των διαστάσεων. Μελετώντας τους πίνακες 3-8, παρατηρείται ότι υπάρχουν σύνολα δεδομένων που η ακρίβεια παραμένει στα ίδια επίπεδα ή ακόμη πολλές φορές αυξάνεται παρά την μείωση των διαστάσεων, και άλλα που η ακρίβεια μειώνεται σε μεγάλο βαθμό.

Στους πίνακες 6 και 8 παρουσιάζονται δύο στήλες αναφορικά με τα κόστη προεπεξεργασίας: Preprocessing cost και Total preprocessing cost. Η πρώτη μέτρηση αφορά το κόστος προεπεξεργασίας που απαιτούν οι αλγόριθμοι (CNN και RSP3) για να δημιουργηθεί το συμπυκνωμένο σύνολο, έχοντας στη διάθεση τους το επεξεργασμένο σύνολο που παρήγαγε ο ENN-rule. Η μέτρηση «Total preprocessing cost» αφορά το συνολικό κόστος προεπεξεργασίας, το οποίο συμπεριλαμβάνει το κόστος προεπεξεργασίας του ENN-rule.

Αναφορικά με την απόδοση των ΤΜΔΕ, παρατηρούμε ότι ο αλγόριθμος CNN-rule επιτυγχάνει υψηλότερα ποσοστά συμπύκνωσης (reduction rate) των χρονοσειρών από ότι ο αλγόριθμος RSP3. Ωστόσο, ο RSP3 επιτυγχάνει υψηλότερη ακρίβεια κατηγοριοποίησης από αυτή του CNN-rule. Όπως ήταν αναμενόμενο, λόγω των αιτιών που παρουσιάστηκαν στην παράγραφο 3.1.1, και οι δύο αλγόριθμοι επιτυγχάνουν υψηλότερα ποσοστά συμπύκνωσης όταν εκτελούνται σε δεδομένα χωρίς θόρυβο. Ωστόσο, η αντίστοιχη μέτρηση της ακρίβειας, παρατηρούμε ότι σε κάποια σύνολα μειώνεται. Τα επίπεδα θορύβου που περιέχεται σε κάθε σύνολο δεδομένων, φαίνεται από στην στήλη «reduction rate» του πίνακα 4. Όσο μεγαλύτερη συμπύκνωση επιτυγχάνει ο αλγόριθμος ENN-rule, τόσο περισσότερο θόρυβο περιέχει.

Λόγω του ότι οι διαφορετικές μέθοδοι αποδίδουν διαφορετικά σε κάθε σύνολο χρονοσειρών, στις επόμενες παραγράφους συγκρίνονται και σχολιάζονται οι αποδόσεις τους για κάθε σύνολο χρονοσειρών χωριστά. Τα ποσοστά που αναγράφονται σε παρενθέσεις κατά τον σχολιασμό, είναι από τους παρακάτω πίνακες 3-8. Στο τέλος επιχειρείται η εξαγωγή γενικών συμπερασμάτων.

| | | | | | 1-NN | |
|---|---|---|---|---|---|---|
| **Dataset** | **Training Set Size** | **Testing Set Size** | **Attributes** | **Classes** | Accuracy | Classification Cost |
| Synthetic Control | 480 | 120 | 60 | 6 | 91,67 | 3.456.000 |
| Synthetic Control 12 | 480 | 120 | 12 | 6 | 98,50 | 691.200 |
| Synthetic Control 6 | 480 | 120 | 6 | 6 | 91,50 | 345.600 |
| Face All | 1800 | 450 | 131 | 14 | 95,07 | 106.110.000 |
| Face All 12 | 1800 | 450 | 12 | 14 | 87,91 | 9.720.000 |
| Face All 6 | 1800 | 450 | 6 | 14 | 68,62 | 4.860.000 |
| Two Patterns | 4000 | 1000 | 128 | 4 | 98,50 | 512.000.000 |
| Two Patterns 12 | 4000 | 1000 | 12 | 4 | 97,56 | 48.000.000 |
| Two Patterns 6 | 4000 | 1000 | 6 | 4 | 80,84 | 24.000.000 |
| Yoga | 2640 | 660 | 426 | 2 | 93,76 | 742.262.400 |
| Yoga 12 | 2640 | 660 | 12 | 2 | 92,36 | 20.908.800 |
| Yoga 6 | 2640 | 660 | 6 | 2 | 88,27 | 10.454.400 |
| Wafer | 5731 | 1433 | 152 | 2 | 99,87 | 1.248.303.496 |
| Wafer 12 | 5731 | 1433 | 12 | 2 | 99,79 | 98.550.276 |
| Wafer 6 | 5731 | 1433 | 6 | 2 | 99,47 | 49.275.138 |
| SwedishLeaf | 900 | 225 | 128 | 15 | 52,36 | 25.920.000 |
| SwedishLeaf 12 | 900 | 225 | 12 | 15 | 52,62 | 2.430.000 |
| SwedishLeaf 6 | 900 | 225 | 6 | 15 | 38,13 | 1.215.000 |
| CBF | 744 | 186 | 128 | 3 | 98,39 | 17.713.152 |
| CBF 12 | 744 | 186 | 12 | 3 | 100,00 | 1.660.608 |
| CBF 6 | 744 | 186 | 6 | 3 | 99,35 | 830.304 |

Πίνακας 3 "Απόδοση κατηγοριοποιητή 1-NN σε μη συμπιεσμένα δεδομένα"

| | | | | | ENN | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Training Set Size** | **Testing Set Size** | **Attributes** | **Classes** | *Preprocessing Cost* | *Items After Processing* | *Reduction Rate* | *Accuracy of new DataSet* | *Classification Cost* |
| **Synthetic Control** | 480 | 120 | 60 | 6 | 6.897.600 | 437,60 | 8,83% | 87,33 | 3.150.720 |
| **Synthetic Control 12** | 480 | 120 | 12 | 6 | 1.379.520 | 471,60 | 1,75% | 98,67 | 679.104 |
| **Synthetic Control 6** | 480 | 120 | 6 | 6 | 689.760 | 418,00 | 12,92% | 91,00 | 300.960 |
| **Face All** | 1800 | 450 | 131 | 14 | 212.102.100 | 1.603,60 | 10,91% | 92,40 | 94.532.220 |
| **Face All 12** | 1800 | 450 | 12 | 14 | 19.429.200 | 1.399,80 | 22,23% | 84,71 | 7.558.920 |
| **Face All 6** | 1800 | 450 | 6 | 14 | 9.714.600 | 991,60 | 44,91% | 68,98 | 2.677.320 |
| **Two Patterns** | 4000 | 1000 | 128 | 4 | 1.023.744.000 | 3.860,80 | 3,48% | 98,12 | 494.182.400 |
| **Two Patterns 12** | 4000 | 1000 | 12 | 4 | 95.976.000 | 3.766,40 | 5,84% | 97,24 | 45.196.800 |
| **Two Patterns 6** | 4000 | 1000 | 6 | 4 | 47.988.000 | 2.826,40 | 29,34% | 83,64 | 16.958.400 |
| **Yoga** | 2640 | 660 | 426 | 2 | 1.483.962.480 | 2.312,00 | 12,42% | 91,76 | 650.041.920 |
| **Yoga 12** | 2640 | 660 | 12 | 2 | 41.801.760 | 2.253,20 | 14,65% | 90,70 | 17.845.344 |
| **Yoga 6** | 2640 | 660 | 6 | 2 | 20.900.880 | 2.101,40 | 20,40% | 87,94 | 8.321.544 |
| **Wafer** | 5731 | 1433 | 152 | 2 | 2.495.735.880 | 5.702,60 | 0,50% | 99,71 | 1.242.117.522 |
| **Wafer 12** | 5731 | 1433 | 12 | 2 | 197.031.780 | 5.706,80 | 0,42% | 99,79 | 98.134.133 |
| **Wafer 6** | 5731 | 1433 | 6 | 2 | 98.515.890 | 5.688,40 | 0,74% | 99,57 | 48.908.863 |
| **SwedishLeaf** | 900 | 225 | 128 | 15 | 51.782.400 | 302,40 | 66,40% | 45,33 | 8.709.120 |
| **SwedishLeaf 12** | 900 | 225 | 12 | 15 | 4.854.600 | 312,20 | 65,31% | 46,93 | 842.940 |
| **SwedishLeaf 6** | 900 | 225 | 6 | 15 | 2.427.300 | 198,40 | 77,96% | 34,67 | 267.840 |
| **CBF** | 744 | 186 | 128 | 3 | 35.378.688 | 725,80 | 2,45% | 98,28 | 17.279.846 |
| **CBF 12** | 744 | 186 | 12 | 3 | 3.316.752 | 744,00 | 0,00% | 100,00 | 1.660.608 |
| **CBF 6** | 744 | 186 | 6 | 3 | 1.658.376 | 732,80 | 1,51% | 99,25 | 817.805 |

**Πίνακας 4 "Απόδοση κατηγοριοποιητή 1-NN σε δεδομένα χωρίς θόρυβο, που προέκυψαν από την επεξεργασία τους με τον αλγόριθμο ENN-rule - Μετρήσεις για την απόδοση του αλγόριθμου ENN-rule"**

| | | | | | CNN | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Training Set Size** | **Testing Set Size** | **Attributes** | **Classes** | *Preprocessing Cost* | *Items After Processing* | *Reduction Rate* | *Accuracy of new DataSet* | *Classification Cost* |
| **Synthetic Control** | 480 | 120 | 60 | 6 | 7.769.448 | 93,60 | 80,50% | 90,17 | 673.920 |
| **Synthetic Control 12** | 480 | 120 | 12 | 6 | 894.070 | 44,40 | 90,75% | 97,00 | 63.936 |
| **Synthetic Control 6** | 480 | 120 | 6 | 6 | 757.608 | 94,20 | 80,38% | 89,50 | 67.824 |
| **Face All** | 1800 | 450 | 131 | 14 | 216.361.382 | 337,00 | 81,28% | 91,60 | 19.866.150 |
| **Face All 12** | 1800 | 450 | 12 | 14 | 30.355.807 | 535,80 | 70,23% | 83,78 | 2.893.320 |
| **Face All 6** | 1800 | 450 | 6 | 14 | 21.466.756 | 903,60 | 49,80% | 64,49 | 2.439.720 |
| **Two Patterns** | 4000 | 1000 | 128 | 4 | 1.169.751.296 | 669,20 | 83,27% | 94,68 | 85.657.600 |
| **Two Patterns 12** | 4000 | 1000 | 12 | 4 | 103.863.223 | 684,60 | 82,89% | 93,52 | 8.215.200 |
| **Two Patterns 6** | 4000 | 1000 | 6 | 4 | 85.572.606 | 1.300,80 | 67,48% | 78,86 | 7.804.800 |
| **Yoga** | 2640 | 660 | 426 | 2 | 1.854.741.378 | 492,80 | 81,33% | 91,58 | 138.555.648 |
| **Yoga 12** | 2640 | 660 | 12 | 2 | 52.225.190 | 556,80 | 78,91% | 90,39 | 4.409.856 |
| **Yoga 6** | 2640 | 660 | 6 | 2 | 30.245.845 | 699,20 | 73,52% | 85,30 | 2.768.832 |
| **Wafer** | 5731 | 1433 | 152 | 2 | 165.882.160 | 62,40 | 98,91% | 99,69 | 13.591.718 |
| **Wafer 12** | 5731 | 1433 | 12 | 2 | 15.627.996 | 70,40 | 98,77% | 99,62 | 1.210.598 |
| **Wafer 6** | 5731 | 1433 | 6 | 2 | 11.503.642 | 104,00 | 98,19% | 99,22 | 894.192 |
| **SwedishLeaf** | 900 | 225 | 128 | 15 | 112.165.197 | 553,40 | 38,51% | 49,87 | 15.937.920 |
| **SwedishLeaf 12** | 900 | 225 | 12 | 15 | 11.331.298 | 569,20 | 36,76% | 49,07 | 1.536.840 |
| **SwedishLeaf 6** | 900 | 225 | 6 | 15 | 4.668.714 | 670,60 | 25,49% | 36,62 | 905.310 |
| **CBF** | 744 | 186 | 128 | 3 | 15.060.198 | 54,00 | 92,74% | 98,17 | 1.285.632 |
| **CBF 12** | 744 | 186 | 12 | 3 | 655.658 | 27,20 | 96,34% | 99,57 | 60.710 |
| **CBF 6** | 744 | 186 | 6 | 3 | 633.654 | 46,00 | 93,82% | 98,17 | 51.336 |

Πίνακας 5 "Απόδοση κατηγοριοποιητή 1-NN σε συμπιεσμένα δεδομένα, που προέκυψαν από την εκτέλεση του αλγορίθμου CNN-rule - Μετρήσεις για την απόδοση του αλγόριθμου CNN-rule"

| | | | | | RSP3 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Training Set Size** | **Testing Set Size** | **Attributes** | **Classes** | *Preprocessing Cost* | *Items After Processing* | *Reduction Rate* | *Accuracy of new DataSet* | *Classification Cost* |
| **Synthetic Control** | 480 | 120 | 60 | 6 | 16.216.416 | 191,60 | 60,08% | 98,33 | 1.379.520 |
| **Synthetic Control 12** | 480 | 120 | 12 | 6 | 3.446.652 | 81,80 | 82,96% | 98,83 | 117.792 |
| **Synthetic Control 6** | 480 | 120 | 6 | 6 | 1.498.098 | 117,60 | 75,50% | 90,83 | 84.672 |
| **Face All** | 1800 | 450 | 131 | 14 | 533.700.917 | 876,20 | 51,32% | 95,46 | 51.651.990 |
| **Face All 12** | 1800 | 450 | 12 | 14 | 50.910.871 | 889,60 | 50,58% | 87,07 | 4.803.840 |
| **Face All 6** | 1800 | 450 | 6 | 14 | 36.037.349 | 1.072,40 | 40,42% | 65,56 | 2.895.480 |
| **Two Patterns** | 4000 | 1000 | 128 | 4 | 2.085.421.107 | 1.902,40 | 52,44% | 98,10 | 243.507.200 |
| **Two Patterns 12** | 4000 | 1000 | 12 | 4 | 196.178.808 | 1.702,00 | 57,45% | 96,66 | 20.424.000 |
| **Two Patterns 6** | 4000 | 1000 | 6 | 4 | 99.448.442 | 1.675,60 | 58,11% | 80,22 | 10.053.600 |
| **Yoga** | 2640 | 660 | 426 | 2 | 4.072.295.880 | 817,40 | 69,04% | 92,85 | 229.820.184 |
| **Yoga 12** | 2640 | 660 | 12 | 2 | 110.559.103 | 847,40 | 67,90% | 91,03 | 6.711.408 |
| **Yoga 6** | 2640 | 660 | 6 | 2 | 51.153.599 | 916,00 | 65,30% | 86,52 | 3.627.360 |
| **Wafer** | 5731 | 1433 | 152 | 2 | 7.196.751.600 | 123,40 | 97,85% | 99,82 | 26.878.494 |
| **Wafer 12** | 5731 | 1433 | 12 | 2 | 495.629.225 | 108,40 | 98,11% | 99,40 | 1.864.046 |
| **Wafer 6** | 5731 | 1433 | 6 | 2 | 207.615.805 | 127,80 | 97,77% | 99,12 | 1.098.824 |
| **SwedishLeaf** | 900 | 225 | 128 | 15 | 1.537.070.054 | 659,80 | 26,69% | 52,00 | 19.002.240 |
| **SwedishLeaf 12** | 900 | 225 | 12 | 15 | 56.002.375 | 659,80 | 26,69% | 51,20 | 1.781.460 |
| **SwedishLeaf 6** | 900 | 225 | 6 | 15 | 21.502.992 | 726,80 | 19,24% | 36,89 | 981.180 |
| **CBF** | 744 | 186 | 128 | 3 | 78.476.032 | 82,80 | 88,87% | 99,78 | 1.971.302 |
| **CBF 12** | 744 | 186 | 12 | 3 | 7.324.284 | 54,80 | 92,63% | 99,68 | 122.314 |
| **CBF 6** | 744 | 186 | 6 | 3 | 3.744.258 | 74,40 | 90,00% | 98,39 | 83.030 |

Πίνακας 6 "Απόδοση κατηγοριοποιητή 1-NN σε συμπιεσμένα δεδομένα, που προέκυψαν από την εκτέλεση του αλγορίθμου RSP3 - Μετρήσεις για την απόδοση του αλγόριθμου RSP3"

| | | | | | ENN-CNN | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Training Set Size | Testing Set Size | Attributes | Classes | Preprocessing Cost | Total Preprocessing Cost | Items After Processing | Reduction Rate | Accuracy of new DataSet | Classification Cost |
| Synthetic Control | 480 | 120 | 60 | 6 | 5.104.296 | 12.001.896 | 68,80 | 85,67% | 85,83 | 495.360 |
| Synthetic Control 12 | 480 | 120 | 12 | 6 | 627.706 | 2.007.226 | 35,80 | 92,54% | 98,00 | 51.552 |
| Synthetic Control 6 | 480 | 120 | 6 | 6 | 343.594 | 1.033.354 | 49,50 | 89,69% | 89,67 | 35.640 |
| Face All | 1800 | 450 | 131 | 14 | 131.258.725 | 343.360.825 | 247,60 | 86,24% | 89,47 | 14.596.020 |
| Face All 12 | 1800 | 450 | 12 | 14 | 11.726.002 | 31.155.202 | 256,80 | 85,73% | 81,07 | 1.386.720 |
| Face All 6 | 1800 | 450 | 6 | 14 | 3.596.684 | 13.311.284 | 218,20 | 87,88% | 65,47 | 589.140 |
| Two Patterns | 4000 | 1000 | 128 | 4 | 942.320.384 | 1.966.064.384 | 607,40 | 84,82% | 93,92 | 77.747.200 |
| Two Patterns 12 | 4000 | 1000 | 12 | 4 | 76.490.842 | 172.466.842 | 543,40 | 86,42% | 93,46 | 6.520.800 |
| Two Patterns 6 | 4000 | 1000 | 6 | 4 | 23.252.021 | 71.240.021 | 413,40 | 89,67% | 81,62 | 2.480.400 |
| Yoga | 2640 | 660 | 426 | 2 | 797.272.206 | 2.281.234.686 | 268,20 | 89,84% | 90,67 | 75.407.112 |
| Yoga 12 | 2640 | 660 | 12 | 2 | 20.693.167 | 62.494.927 | 268,00 | 89,85% | 89,67 | 2.122.560 |
| Yoga 6 | 2640 | 660 | 6 | 2 | 10.042.098 | 30.942.978 | 248,80 | 90,58% | 86,94 | 985.248 |
| Wafer | 5731 | 1433 | 152 | 2 | 93.401.720 | 2.589.137.600 | 40,00 | 99,30% | 99,61 | 8.712.640 |
| Wafer 12 | 5731 | 1433 | 12 | 2 | 8.257.361 | 205.289.141 | 42,80 | 99,25% | 99,60 | 735.989 |
| Wafer 6 | 5731 | 1433 | 6 | 2 | 4.111.356 | 102.627.246 | 46,60 | 99,19% | 99,47 | 400.667 |
| SwedishLeaf | 900 | 225 | 128 | 15 | 8.253.005 | 60.035.405 | 98,60 | 89,04% | 43,29 | 2.839.680 |
| SwedishLeaf 12 | 900 | 225 | 12 | 15 | 792.067 | 5.646.667 | 104,20 | 88,42% | 45,60 | 281.340 |
| SwedishLeaf 6 | 900 | 225 | 6 | 15 | 187.600 | 2.614.900 | 75,40 | 91,62% | 34,76 | 101.790 |
| CBF | 744 | 186 | 128 | 3 | 11.888.691 | 47.267.379 | 47,20 | 93,66% | 97,63 | 1.123.738 |
| CBF 12 | 744 | 186 | 12 | 3 | 655.658 | 3.972.410 | 27,20 | 96,34% | 99,57 | 60.710 |
| CBF 6 | 744 | 186 | 6 | 3 | 513.030 | 2.171.406 | 37,20 | 95,00% | 98,06 | 41.515 |

Πίνακας 7 "Απόδοση κατηγοριοποιητή 1-NN σε συμπιεσμένα δεδομένα, που προέκυψαν από τη διαδοχική εκτέλεση των αλγορίθμων ENN-rule και CNN-rule - Μετρήσεις σχετικά με την απόδοση του αλγόριθμου CNN-rule σε δεδομένα χωρίς θόρυβο, που προέκυψαν από την επεξεργασία τους με τον αλγόριθμο ENN-rule"

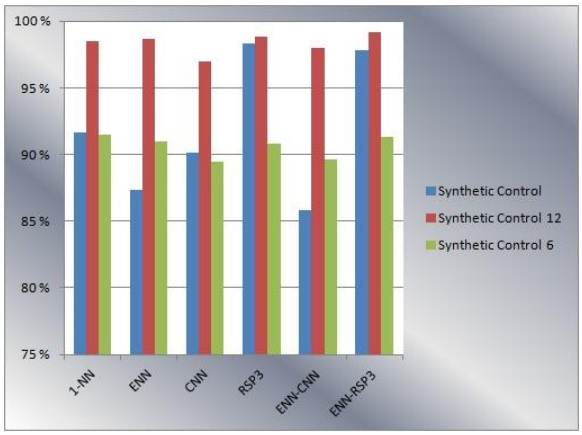
| | | | | | ENN-RSP3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Training Set Size | Testing Set Size | Attributes | Classes | Preprocessing Cost | Total Preprocessing Cost | Items After Processing | Reduction Rate | Accuracy of new DataSet | Classification Cost |
| Synthetic Control | 480 | 120 | 60 | 6 | 13.809.252 | 20.706.852 | 157,00 | 67,29% | 97,83 | 1.130.400 |
| Synthetic Control 12 | 480 | 120 | 12 | 6 | 3.331.934 | 4.711.454 | 77,20 | 83,92% | 99,17 | 111.168 |
| Synthetic Control 6 | 480 | 120 | 6 | 6 | 1.148.976 | 1.838.736 | 79,00 | 83,54% | 91,33 | 56.880 |
| Face All | 1800 | 450 | 131 | 14 | 433.892.174 | 645.994.274 | 742,40 | 58,76% | 92,58 | 43.764.480 |
| Face All 12 | 1800 | 450 | 12 | 14 | 30.208.452 | 49.637.652 | 551,60 | 69,36% | 84,27 | 2.978.640 |
| Face All 6 | 1800 | 450 | 6 | 14 | 9.131.255 | 18.845.855 | 364,00 | 79,78% | 67,33 | 982.800 |
| Two Patterns | 4000 | 1000 | 128 | 4 | 1.940.202.957 | 2.963.946.957 | 1.827,20 | 54,32% | 97,54 | 233.881.600 |
| Two Patterns 12 | 4000 | 1000 | 12 | 4 | 173.979.535 | 269.955.535 | 1.568,60 | 60,79% | 96,36 | 18.823.200 |
| Two Patterns 6 | 4000 | 1000 | 6 | 4 | 49.989.022 | 97.977.022 | 754,60 | 81,14% | 81,90 | 4.527.600 |
| Yoga | 2640 | 660 | 426 | 2 | 3.070.972.230 | 4.554.934.710 | 524,60 | 80,13% | 91,70 | 147.496.536 |
| Yoga 12 | 2640 | 660 | 12 | 2 | 77.006.849 | 118.808.609 | 482,20 | 81,73% | 90,24 | 3.819.024 |
| Yoga 6 | 2640 | 660 | 6 | 2 | 32.935.853 | 53.836.733 | 413,20 | 84,35% | 87,45 | 1.636.272 |
| Wafer | 5731 | 1433 | 152 | 2 | 7.049.709.688 | 9.545.445.568 | 96,60 | 98,31% | 99,64 | 21.041.026 |
| Wafer 12 | 5731 | 1433 | 12 | 2 | 491.500.121 | 688.531.901 | 79,00 | 98,62% | 99,48 | 1.358.484 |
| Wafer 6 | 5731 | 1433 | 6 | 2 | 200.314.802 | 298.830.692 | 72,00 | 98,74% | 99,36 | 619.056 |
| SwedishLeaf | 900 | 225 | 128 | 15 | 119.553.792 | 171.336.192 | 158,40 | 82,40% | 45,16 | 4.561.920 |
| SwedishLeaf 12 | 900 | 225 | 12 | 15 | 7.434.763 | 12.289.363 | 162,00 | 82,00% | 46,40 | 437.400 |
| SwedishLeaf 6 | 900 | 225 | 6 | 15 | 991.218 | 3.418.518 | 102,80 | 88,58% | 32,98 | 138.780 |
| CBF | 744 | 186 | 128 | 3 | 74.427.750 | 109.806.438 | 78,00 | 89,52% | 99,68 | 1.857.024 |
| CBF 12 | 744 | 186 | 12 | 3 | 7.324.284 | 10.641.036 | 54,80 | 92,63% | 99,68 | 122.314 |
| CBF 6 | 744 | 186 | 6 | 3 | 3.624.030 | 5.282.406 | 66,00 | 91,13% | 98,71 | 73.656 |

Πίνακας 8 "Απόδοση κατηγοριοποιητή 1-NN σε συμπιεσμένα δεδομένα, που προέκυψαν από τη διαδοχική εκτέλεση των αλγορίθμων ENN-rule και RSP3 - Μετρήσεις σχετικά με την απόδοση του αλγόριθμου RSP3 σε δεδομένα χωρίς θόρυβο, που προέκυψαν από την επεξεργασία τους με τον αλγόριθμο ENN-rule"

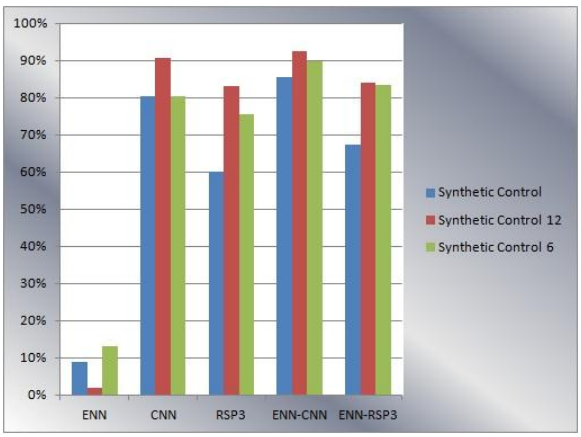
### 4.2.1 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ SYNTETIC CONTROL

Πριν το σχολιασμό των πειραμάτων, πρέπει να διευκρινιστεί ότι οι μετρήσεις του κόστους προεπεξεργασίας (διάγραμμα (δ) στα σχήματα 16-22), που αφορούν τις μεθόδους ENN-CNN και ENN-RSP3, συμπεριλαμβάνουν και το κόστος του ENN. Οι μετρήσεις αυτές αφορούν τα κόστη των αλγορίθμων CNN και RSP3, που εκτελούνται στα επεξεργασμένα από τον ENN-rule δεδομένα. Το συνολικό κόστος προεπεξεργασίας για τις προαναφερθείσες μεθόδους προκύπτει αθροίζοντας και τις αντίστοιχες μετρήσεις κόστους που αφορούν τον αλγόριθμο ENN-rule. Το κόστος του ENN-rule, όπως έχει ήδη αναφερθεί σε προηγούμενο κεφάλαιο (βλέπε παράγραφο 3.1.5) είναι ίσο με $\frac{N \cdot (N-1)}{2}$, όπου *N* είναι το πλήθος των αντικειμένων του συνόλου δεδομένων εκπαίδευσης.

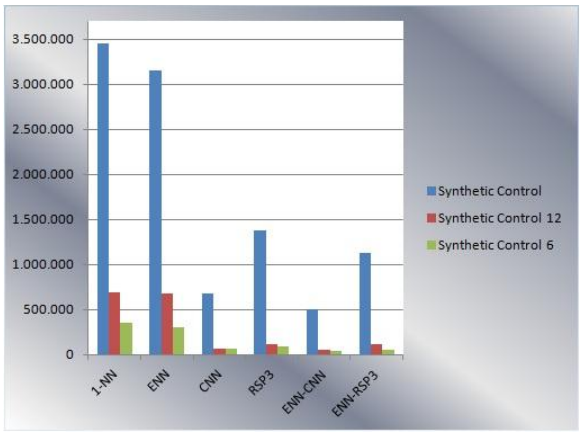
Στο σύνολο δεδομένων Synthetic Control παρατηρούμε ένα φαινόμενο που δεν συναντάται συχνά. Το συμπυκνωμένο σύνολο που προκύπτει από τον αλγόριθμο RSP3, βελτιώνει την ακρίβεια (Σχήμα 16α) του κατηγοριοποιητή 1-NN (98,33%) σε σχέση με το αρχικό σύνολο δεδομένων (91,67%) και παράλληλα συμπιέζει σε σημαντικό βαθμό τα δεδομένα (Σχήμα 16β, 60,08%). Το σύνολο αυτό περιέχει ελάχιστο θόρυβο. Έτσι, τα ποσοστά μείωσης του ENN-rule είναι σχετικά χαμηλά (8,83%). Συνεπώς, οι μετρήσεις ακρίβειας και ποσοστών μείωσης των αλγορίθμων ENN-CNN (85,67%) και ENN-RSP3 (97,83%), διαφέρουν λίγο με τους αντίστοιχους CNN (90,17%) και RSP3 (98,33%). Κάτι που αξίζει να σημειωθεί, είναι ότι η τεχνική PAA παρήγαγε ένα σύνολο δώδεκα διαστάσεων, το οποίο σε κάθε περίπτωση βελτιώνει την ακρίβεια (98,83%) και τα ποσοστά μείωσης (Πίνακες 5-8, στήλη 'Reduction Rate') όλων των αλγορίθμων. Επιπρόσθετα, συμπυκνωμένο σύνολο που παρήγαγε ο αλγόριθμος RSP3 επιτυγχάνει υψηλότερη ακρίβεια (90,33%) από ότι αυτό του CNN-rule (90,17%), ενώ ο CNN-rule επιτυγχάνει υψηλότερα ποσοστά μείωσης (80,50% έναντι 60,08% του RSP3). Τέλος, παρατηρούμε ότι το κόστος προεπεξεργασίας για του CNN-rule (7,8 εκατ. υπολογισμοί), είναι πολύ χαμηλότερο από αυτό του RSP3 (16,2 εκατ. υπολογισμοί). Περισσότερες λεπτομέρειες φαίνονται και στο σχήμα 16.



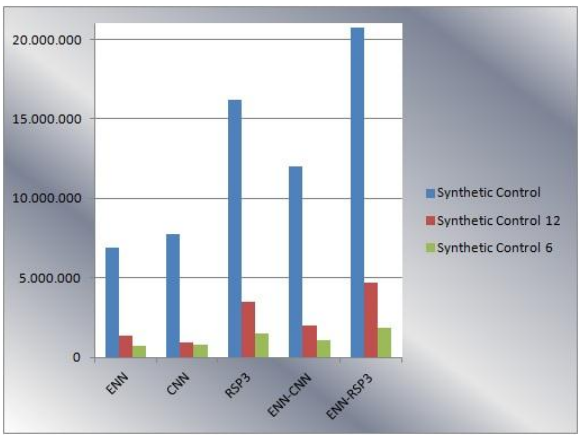

(α) ακρίβεια κατηγοριοποίησης




(β) ποσοστό μείωσης




(γ) κόστος κατηγοριοποίησης



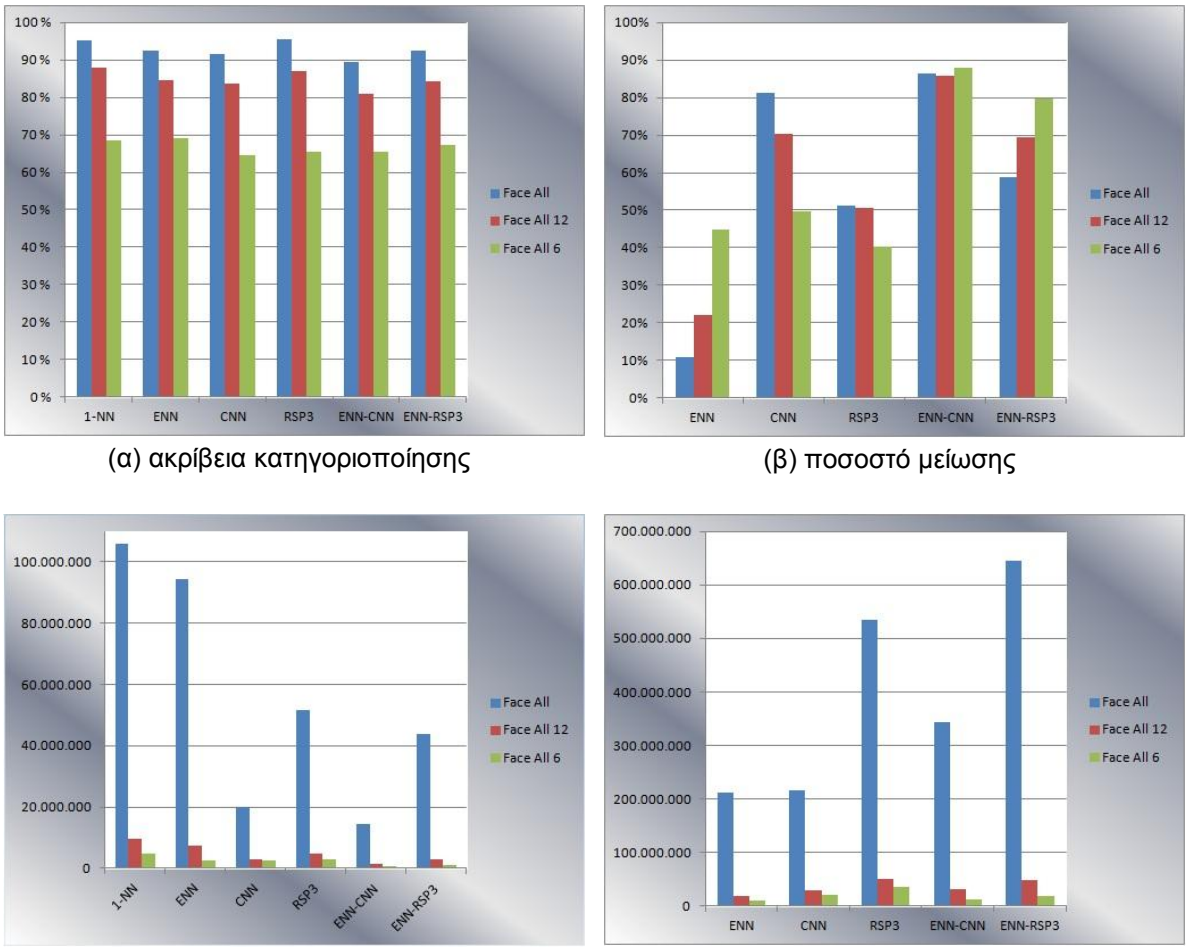

(δ) κόστος προεπεξεργασίας

**Σχήμα 16 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων Synthetic Control"**

### 4.2.2 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ FACE ALL

Το σύνολο δεδομένων Face All, καθώς επίσης και όλα τα συμπυκνωμένα σύνολα δεδομένων που προέκυψαν με την επεξεργασία των αλγορίθμων, επιτυγχάνουν υψηλή ακρίβεια κατηγοριοποίησης (Σχήμα 17α). Ο αλγόριθμος RSP3, όπως και στο σύνολο δεδομένων Synthetic Control, καταφέρνει για ακόμη μια φορά να βελτιώσει την ακρίβεια του κατηγοριοποιητή 1-NN (95,46%), έστω και οριακά (95,07%). Δεν καταφέρνει όμως να συμπιέσει τα δεδομένα σε τόσο μεγάλο βαθμό (51,32%). Αντίθετα οι υπόλοιποι αλγόριθμοι επιτυγχάνουν ένα ικανοποιητικό βαθμό μείωσης (Πίνακες 5-8, στήλη 'Reduction Rate'). Εξαιτίας του αυξημένου όγκου δεδομένων και των πολλών διαστάσεων (131 χαρακτηριστικά) του συγκεκριμένου συνόλου, το κόστος κατηγοριοποίησης (Σχήμα 17γ) καθώς και το κόστος προεπεξεργασίας (Σχήμα 17δ) είναι αρκετά υψηλό. Αυτός είναι και ο λόγος που η μείωση σε δώδεκα διαστάσεις μπορεί να θεωρηθεί επιτυχημένη, αφού η ακρίβεια κατηγοριοποίησης όλων των μεθόδων μειώνεται ελάχιστα (Σχήμα 17α), ενώ παράλληλα τα κόστη κατηγοριοποίησης (Σχήμα 17γ) και

προεπεξεργασίας (Σχήμα 17δ) ελαχιστοποιούνται. Οι μετρήσεις ακρίβειας των συμπυκνωμένων συνόλων που προέκυψαν με τις μεθόδους CNN και RSP3 (και των αντίστοιχων ENN-CNN και ENN-RSP3) δεν διαφέρουν σημαντικά (Πίνακες 5-8). Ωστόσο, ο CNN επιτυγχάνει μεγαλύτερη συμπύκνωση (81,28%) και ως εκ τούτου μικρότερο κόστος κατηγοριοποίησης (19,9 εκατ. υπολογισμοί), καθώς επίσης έχει και χαμηλότερο κόστος προεπεξεργασίας (216,4 εκατ. υπολογισμοί) σε σύγκριση με τις αντίστοιχες μετρήσεις που αφορούν τον RSP3 (51,32% - 51,7 εκατ. υπολογισμοί - 533,7 εκατ. υπολογισμοί αντίστοιχα). Οι διαφορές που προαναφέρθηκαν μπορούν να εντοπιστούν εύκολα και στο σχήμα 17.




(α) ακρίβεια κατηγοριοποίησης

(β) ποσοστό μείωσης

(γ) κόστος κατηγοριοποίησης

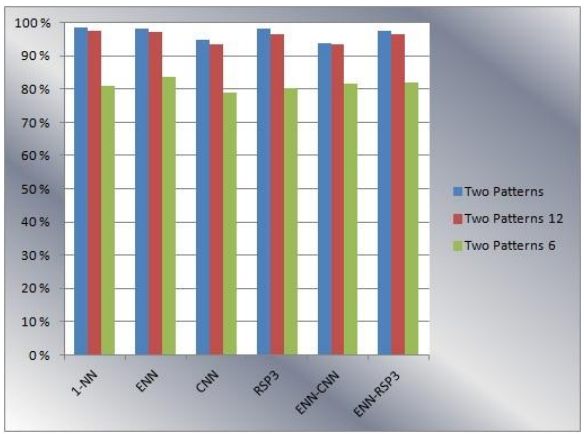
(δ) κόστος προεπεξεργασίας

**Σχήμα 17 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων Face All"**

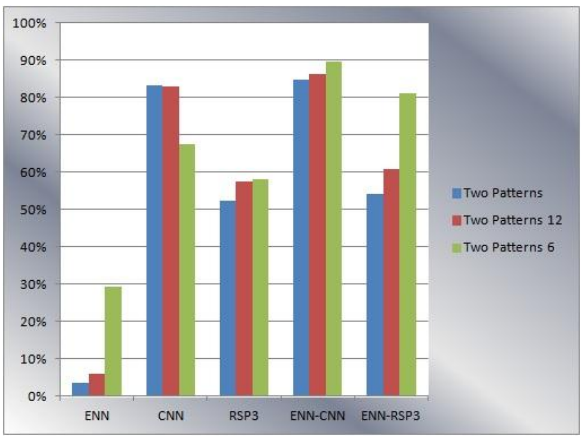
### 4.2.3 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ TWO PATTERNS

Ανάλογα με το σύνολο δεδομένων Face All, το Two Patterns καθώς και όλα τα συμπυκνωμένα σύνολα δεδομένων, που προέκυψαν με την επεξεργασία των αλγορίθμων, επιτυγχάνουν πολύ υψηλή ακρίβεια κατηγοριοποίησης (Σχήμα 18α).

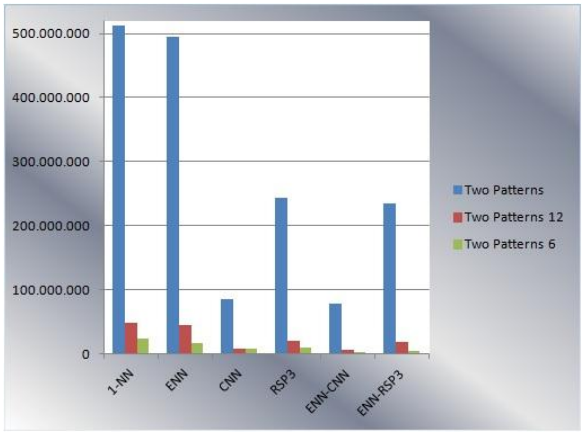
Οι αλγόριθμοι ENN και RSP3 καταφέρνουν δημιουργήσουν σύνολα δεδομένων που διατηρούν την ακρίβεια στα ίδια επίπεδα (98,12% και 98,10% αντίστοιχα) με αυτή του αρχικού μη συμπυκνωμένου συνόλου (98,50%). Ο CNN παρουσιάζει αυξημένο ποσοστό μείωσης του όγκου των δεδομένων (83,27%) και παρόλα αυτά διατηρείτε η ακρίβεια κατηγοριοποίησης σε αυξημένα επίπεδα (94,68%). Οι υπόλοιποι αλγόριθμοι επιτυγχάνουν και αυτοί ένα ικανοποιητικό βαθμό μείωσης (Πίνακες 4-8, στήλη 'Reduction Rate'). Λόγο του πολύ μεγάλου όγκου δεδομένων (5000 αντικείμενα) και των πολλών διαστάσεων (128 χαρακτηριστικά) του συγκεκριμένου συνόλου, το κόστος κατηγοριοποίησης (512 εκατ. υπολογισμοί) καθώς και το κόστος προεπεξεργασίας (Πίνακες 4-8, στήλη 'Preprocessing Cost') είναι ιδιαίτερα υψηλό. Αυτός είναι και ο λόγος που μείωση σε δώδεκα διαστάσεις μπορεί να θεωρηθεί επιτυχημένη, αφού η ακρίβεια κατηγοριοποίησης όλων των μεθόδων μειώνεται ελάχιστα (Πίνακες 3-8, στήλες 'Accuracy' και 'Accuracy of new DataSet'), ενώ παράλληλα τα κόστη κατηγοριοποίησης και προεπεξεργασίας ελαχιστοποιούνται (Πίνακες 3-8, στήλες 'Classification Cost' και 'Preprocessing Cost' αντίστοιχα). Και σε αυτή την περίπτωση, οι μετρήσεις ακρίβειας κατηγοριοποίησης της διαδικασίας συμπίεσης με τις μεθόδους CNN και RSP3 (και των αντίστοιχων ENN-CNN και ENN-RSP3) δεν διαφέρουν σημαντικά (Πίνακες 5-8, στήλη 'Accuracy of new DataSet'). Ο CNN, και ο ENN-CNN κατ' επέκταση, επιτυγχάνουν πάλι πολύ μεγάλη συμπύκνωση (82,89% και 86,42% αντίστοιχα) και ως εκ τούτου πολύ μικρό κόστος κατηγοριοποίησης (Σχήμα 18γ). Επιπρόσθετα έχουν χαμηλότερο κόστος προεπεξεργασίας σε σύγκριση με τις αντίστοιχες μετρήσεις που αφορούν τον RSP3 και τον ENN-RSP3 (Σχήμα 18δ).



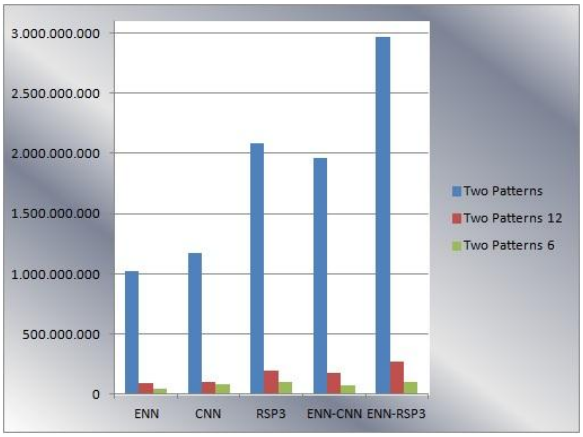

(α) ακρίβεια κατηγοριοποίησης




(β) ποσοστό μείωσης




(γ) κόστος κατηγοριοποίησης

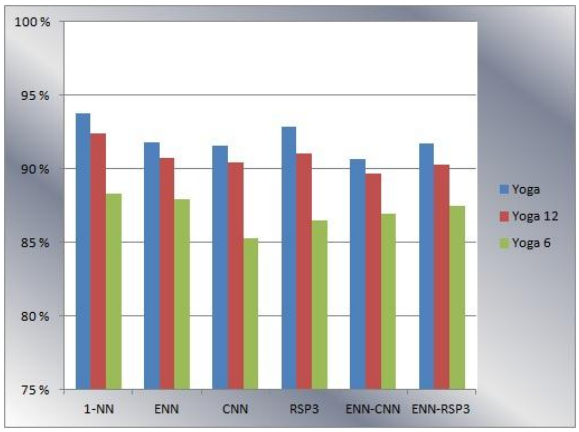



(δ) κόστος προεπεξεργασίας

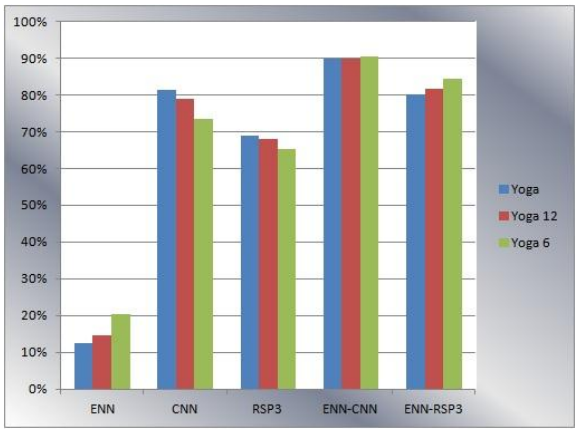
**Σχήμα 18 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων Two Patterns"**

### 4.2.4 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ YOGA

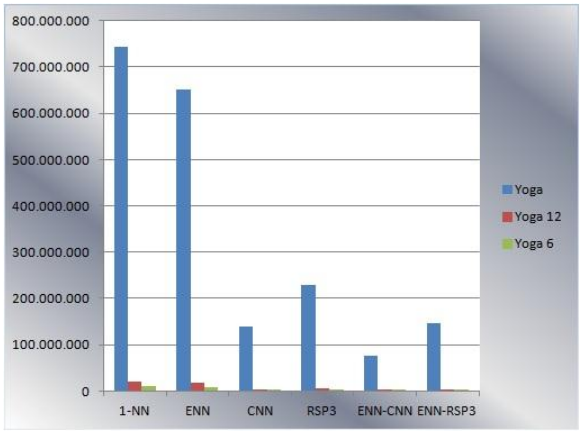
Το σύνολο δεδομένων Yoga χαρακτηρίζεται από τον πολύ μεγάλο αριθμό των διαστάσεων (426). Όλα τα συμπυκνωμένα σύνολα δεδομένων, που προκύπτουν με την επεξεργασία των αλγορίθμων, επιτυγχάνουν πολύ υψηλή ακρίβεια κατηγοριοποίησης (Σχήμα 19α). Επίσης, η τεχνική PAA, παρήγαγε ένα σύνολο δώδεκα διαστάσεων που επηρεάζει ελάχιστα την ακρίβεια και τα ποσοστά μείωσης όλων των αλγορίθμων (Σχήμα 19α και 19β). Λόγο του μεγάλου όγκου δεδομένων και των πολλών διαστάσεων του συγκεκριμένου συνόλου, το κόστος κατηγοριοποίησης (742 εκατ. υπολογισμοί) καθώς και το κόστος προεπεξεργασίας είναι ιδιαίτερα υψηλό (Πίνακες 4-8, στήλη 'Preprocessing Cost'). Αυτός είναι και ο λόγος που μείωση σε δώδεκα διαστάσεις μπορεί να θεωρηθεί επιτυχημένη, αφού η ακρίβεια όλων των μεθόδων μειώνεται ελάχιστα, ενώ παράλληλα τα κόστη κατηγοριοποίησης και προεπεξεργασίας ελαχιστοποιούνται (Πίνακες 3-8).



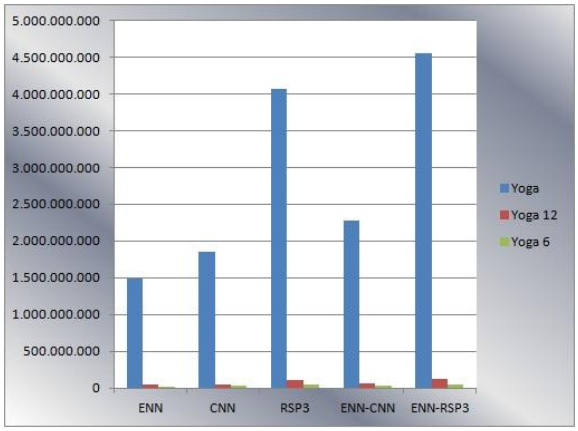

(α) ακρίβεια κατηγοριοποίησης




(β) ποσοστό μείωσης




(γ) κόστος κατηγοριοποίησης

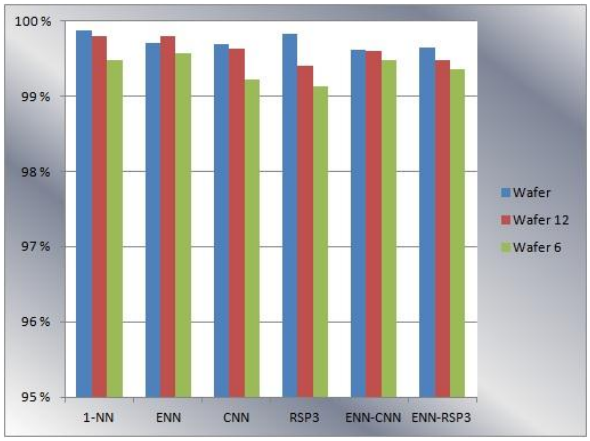



(δ) κόστος προεπεξεργασίας

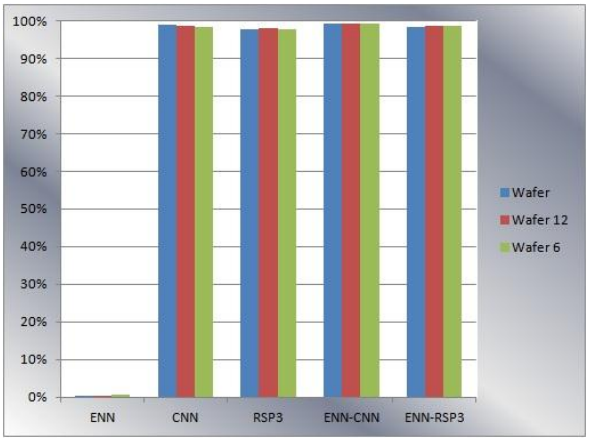
**Σχήμα 19 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων Yoga"**

### 4.2.5 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ WAFER

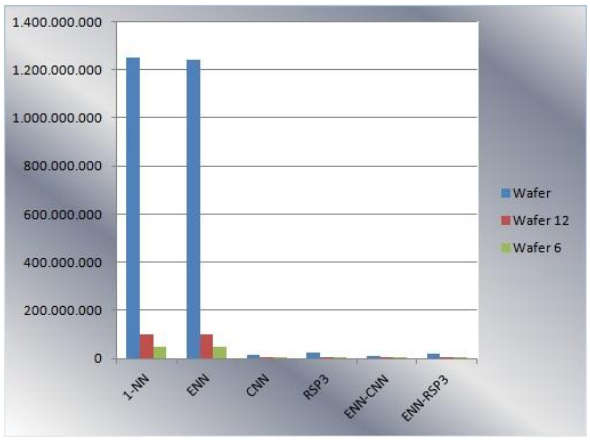
Το σύνολο δεδομένων Wafer χαρακτηρίζεται από τα ελάχιστα επίπεδα θορύβου που εμπεριέχει. Αυτό είναι εμφανές από τα ιδιαίτερα χαμηλά ποσοστά μείωσης του ENN-rule (0,50%). Ως αποτέλεσμα, δεν εντοπίζονται ιδιαίτερες διαφορές μεταξύ των αλγορίθμων που εκτελούνται σε δεδομένα επεξεργασμένα από τον ENN-rule ή όχι (Πίνακες 3-8). Όλες οι μέθοδοι μπορούν να επιτύχουν ιδιαίτερα υψηλά επίπεδα ακρίβειας κατηγοριοποίησης (>99%) και υψηλά ποσοστά συμπίεσης (>97,85%). Επίσης, η μείωση των χρονοσειρών σε δώδεκα και έξι διαστάσεις μπορεί να θεωρηθεί επιτυχημένη αφού η ακρίβεια δεν πέφτει κάτω από το 99% ενώ παράλληλα τα κόστη ελαχιστοποιούνται (Σχήμα 20γ και 20δ).



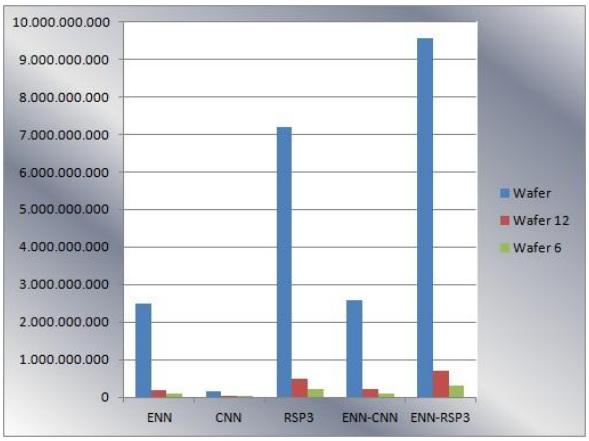

(α) ακρίβεια κατηγοριοποίησης




(β) ποσοστό μείωσης




(γ) κόστος κατηγοριοποίησης



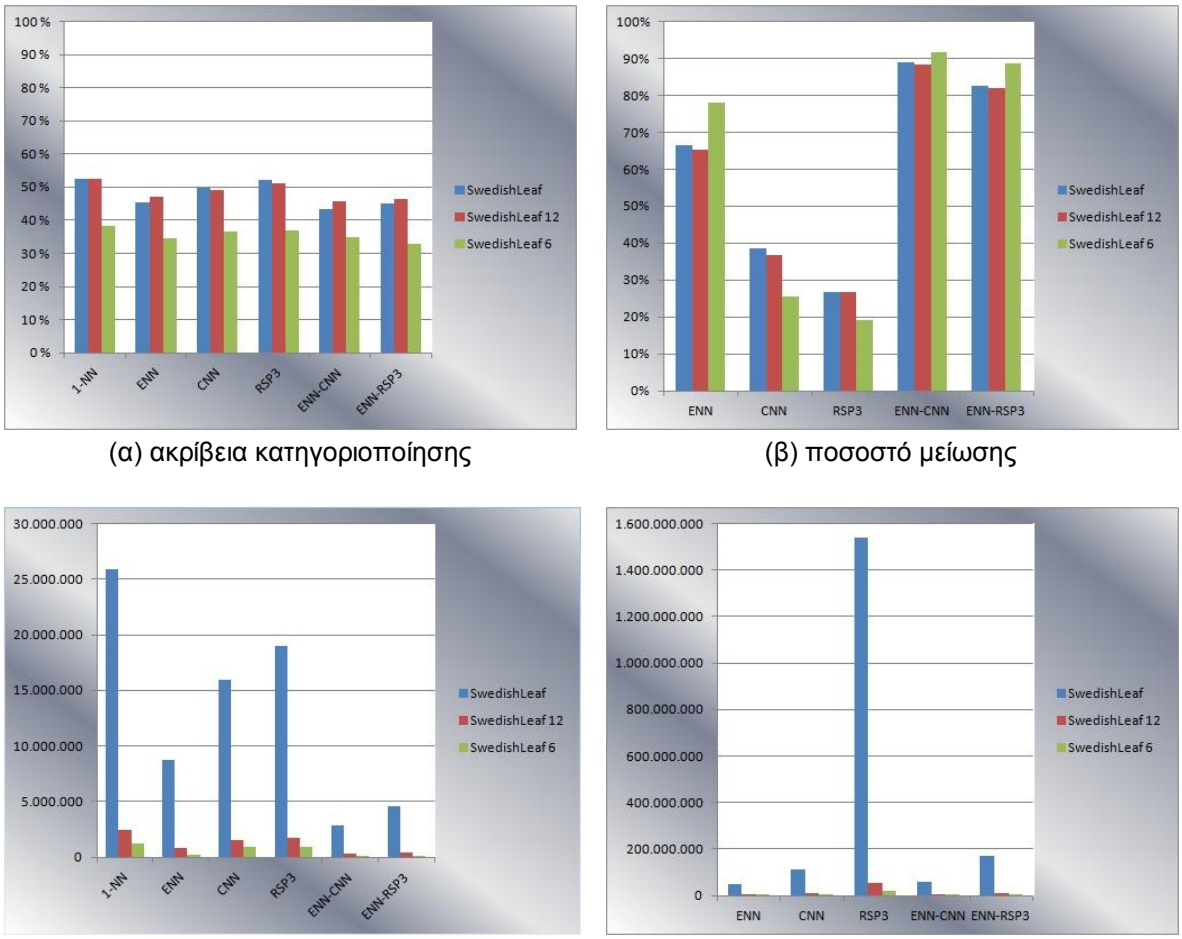

(δ) κόστος προεπεξεργασίας

**Σχήμα 20 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων Wafer"**

### 4.2.6 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ SWEDISH LEAF

Στο σύνολο δεδομένων Swedish leaf, όλοι οι αλγόριθμοι δεν καταφέρνουν να επιτύχουν υψηλή ακρίβεια κατηγοριοποίησης (<52%). Αν και ο αλγόριθμος ENN-rule θεωρεί θόρυβο μεγάλο ποσοστό (66,40%) των χρονοσειρών εκπαίδευσης και τις απομακρύνει, ο κατηγοριοποιητής 1-NN, που χρησιμοποιεί το σύνολο δεδομένων που παράγει ο ENN-rule, δε μπορεί να επιτύχει υψηλότερη ακρίβεια (45,33%). Αυτό έχει αντίκτυπο και στους αλγορίθμους ENN-CNN και ENN-RSP3 (Πίνακες 7-8). Και εδώ παρατηρούμε ότι οι μείωση σε δώδεκα διαστάσεις μπορεί να θεωρηθεί επιτυχημένη, αφού η ακρίβεια όλων των μεθόδων δεν μειώνεται (Σχήμα 21α), ενώ παράλληλα τα κόστη κατηγοριοποίησης και προεπεξεργασίας ελαχιστοποιούνται (Σχήμα 21γ και 21δ). Οι μετρήσεις της ακρίβειας κατηγοριοποίησης των μεθόδων CNN και RSP3 (και των αντίστοιχων ENN-CNN και ENN-RSP3) δεν διαφέρουν σημαντικά (Πίνακες 5-8, στήλη 'Accuracy of new DataSet'). Ωστόσο, ο CNN επιτυγχάνει μεγαλύτερη συμπύκνωση (38,51%) και ως εκ τούτου μικρότερο κόστος κατηγοριοποίησης (15,9 εκατ. υπολογισμοί) και

κόστος προεπεξεργασίας (112,2 εκατ. υπολογισμοί) σε σύγκριση με τις αντίστοιχες μετρήσεις που αφορούν τον RSP3 (26,69% - 19 εκατ. υπολογισμοί - 1.537 εκατ. υπολογισμοί αντίστοιχα).




(α) ακρίβεια κατηγοριοποίησης

(β) ποσοστό μείωσης

(γ) κόστος κατηγοριοποίησης

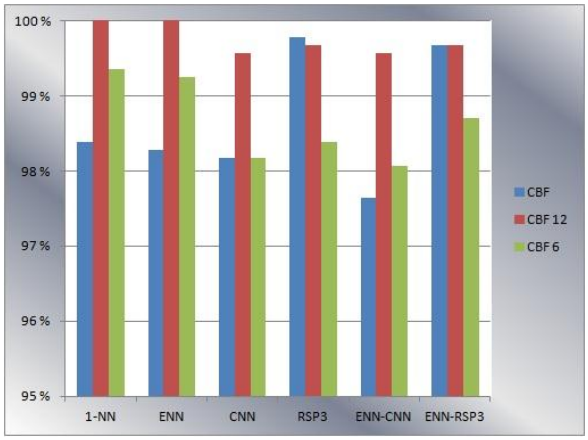
(δ) κόστος προεπεξεργασίας

**Σχήμα 21 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων Swedish Leaf"**

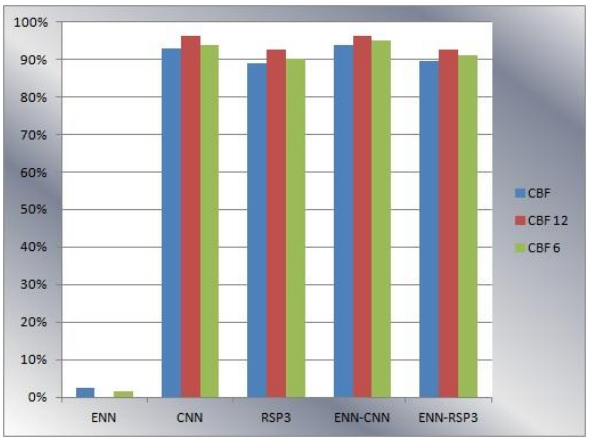
### 4.2.7 ΣΥΝΟΛΟ ΧΡΟΝΟΣΕΙΡΩΝ CBF

Ακόμη μια περίπτωση συνόλου χρονοσειρών όπου το συμπυκνωμένο σύνολο δεδομένων που δημιουργεί ο αλγόριθμος RSP3, σε σχέση με τις πραγματικές διαστάσεις του συνόλου δεδομένων, μπορεί να βελτιώσει (99,78%) την ακρίβεια του κατηγοριοποιητή 1-NN (98,39%) και παράλληλα να συμπιέσει σε σημαντικό βαθμό τα δεδομένα (88,87%). Το σύνολο αυτό, όπως και το σύνολο wafer, περιέχει ελάχιστο θόρυβο. Έτσι, τα ποσοστά μείωσης του ENN-rule είναι ιδιαίτερα χαμηλά (2,45%). Συνεπώς, οι μετρήσεις ακρίβειας κατηγοριοποίησης και ποσοστών μείωσης των αλγορίθμων ENN-CNN και ENN-RSP3 (93,66% και

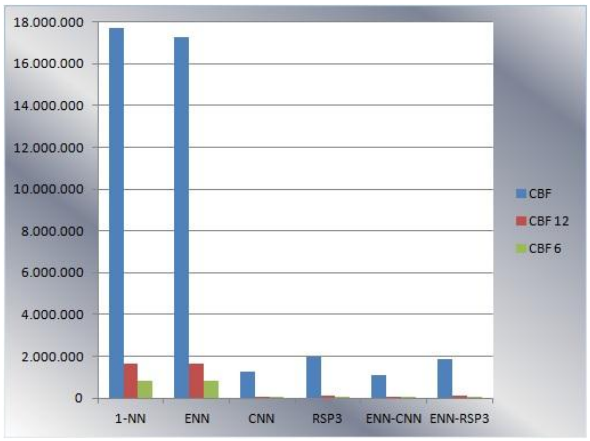
89,52%) δεν διαφέρουν σημαντικά με τους αντίστοιχους CNN και RSP3 (92,74% και 88,87%). Επίσης το κόστος κατηγοριοποίησης και των τεσσάρων μεθόδων ελαχιστοποιείται σε σχέση με αυτό του 1-NN (Σχήμα 22γ). Η τεχνική PAA παρήγαγε ένα σύνολο δώδεκα διαστάσεων που δεν επηρεάζει αρνητικά την ακρίβεια κατηγοριοποίησης και τα ποσοστά μείωσης όλων των αλγορίθμων (Πίνακες 3-8). Και σε αυτή την περίπτωση, η πειραματική διαδικασία με τον αλγόριθμο RSP3 επιτυγχάνει υψηλότερη ακρίβεια (99,68%) από ότι αυτή με τον CNN-rule (99,57%), ενώ ο CNN-rule επιτυγχάνει υψηλότερα (96,34%) ποσοστά μείωσης (92,63%). Κάτι που χρήζει αναφοράς, είναι ότι το σύνολο δεδομένων με τις δώδεκα διαστάσεις, στις περισσότερες περιπτώσεις (πλην αυτής του RSP3), βελτιώνει την ακρίβεια κατηγοριοποίησης σε σχέση με αυτή του αρχικού συνόλου. Τέλος παρατηρούμε ότι το κόστος προεπεξεργασίας του CNN-rule, είναι πολύ χαμηλότερο από αυτό του RSP3 (Σχήμα 22δ).



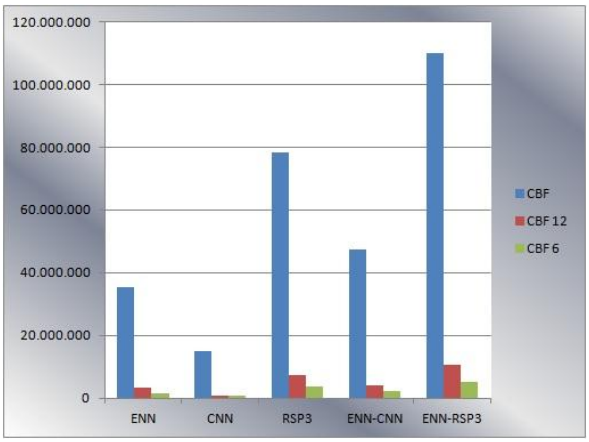

(α) ακρίβεια κατηγοριοποίησης




(β) ποσοστό μείωσης




(γ) κόστος κατηγοριοποίησης




(δ) κόστος προεπεξεργασίας

**Σχήμα 22 "Μετρήσεις απόδοσης μεθόδων στο σύνολο δεδομένων CBF"**

### 4.2.8 ΣΥΜΠΕΡΑΣΜΑΤΑ ΠΕΙΡΑΜΑΤΩΝ

Μελετώντας τις μετρήσεις της παρούσας πειραματικής μελέτης, η εφαρμογή παραδοσιακών αλγορίθμων μείωσης του όγκου των δεδομένων σε δεδομένα χρονοσειρών θεωρείται αποτελεσματική. Συγκεκριμένα, προκύπτουν τα ακόλουθα συμπεράσματα αναφορικά με τα συγκεκριμένα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική μελέτη:

- Η μείωση των διαστάσεων μπορεί να επιταχύνει σε μεγάλο βαθμό την διαδικασία κατηγοριοποίησης χρονοσειρών. Ωστόσο, η ακρίβεια δεν διατηρείται πάντα σε υψηλά επίπεδα. Η μείωση στις δώδεκα διαστάσεις, στα περισσότερα σύνολα χρονοσειρών, κρίνεται ικανοποιητική. Αυτό δεν ισχύει για την μείωσή στις έξι διαστάσεις.
- Για την επίτευξη υψηλών ποσοστών συμπύκνωσης στα σύνολα δεδομένων που περιέχουν υψηλά επίπεδα θορύβου είναι απαραίτητη η απομάκρυνση του θορύβου. Ωστόσο, τα πειραματικά αποτελέσματα αποδεικνύουν ότι αυτό δεν βοηθάει την αύξηση της ακρίβειας. Από την άλλη, η εκτέλεση ενός αλγορίθμου επεξεργασίας, όπως ο ENN-rule, εισάγει επιπρόσθετο κόστος προεπεξεργασίας που ίσως είναι περιττό στις περιπτώσεις που τα σύνολα έχουν χαμηλά ως ανύπαρκτα επίπεδα θορύβου.
- Ο αλγόριθμος σύνοψης RSP3 είναι σε θέση ακόμη και να αυξήσει την ακρίβεια του κατηγοριοποιητή εγγύτερων γειτόνων. Το ίδιο δεν ισχύει για τον αλγόριθμο επιλογής CNN-rule.
- Ο αλγόριθμος CNN-rule επιτυγχάνει υψηλότερα ποσοστά μείωσης των χρονοσειρών από ότι ο αλγόριθμος RSP3.
- Ο αλγόριθμος CNN-rule απαιτεί πολύ χαμηλότερο κόστος προεπεξεργασίας σε σχέση με τον RSP3.
- Ο αλγόριθμος RSP3 επιτυγχάνει συνήθως υψηλότερη ακρίβεια κατηγοριοποίησης σε σχέση με τον αλγόριθμο CNN-rule
- Δεν είναι δυνατό να αξιολογηθεί μια μέθοδος καλύτερη από την άλλη. Αυτό εξαρτάται από το είδος της εφαρμογής. Αν βασικό κριτήριο είναι η γρήγορη κατηγοριοποίηση και το χαμηλό κόστος προεπεξεργασίας, ο αλγόριθμος CNN-rule φαίνεται να είναι καλύτερος. Από την άλλη, αν βασικότερο κριτήριο είναι η υψηλή ακρίβεια ενώ η ταχύτητα θεωρείται δευτερεύουσας σημασίας, ο RSP3 είναι ιδιαίτερα αποτελεσματικός αφού είναι σε θέση ακόμη και να επιτύχει βελτιώσεις προς αυτή την κατεύθυνση.

## ΕΠΙΛΟΓΟΣ

Στο κεφάλαιο παρουσιάστηκε μια πειραματική μελέτη που παρουσίασε και σύγκρινε τις μετρήσεις απόδοσης που προέκυψαν από την εκτέλεση

παραδοσιακών, μη παραμετρικών τεχνικών μείωσης όγκου δεδομένων σε σύνολα χρονοσειρών. Τέτοιου είδους εφαρμογή θεωρείται κάτι το πρωτότυπο αφού δεν το συναντάμε στην διεθνή βιβλιογραφία. Μελετώντας τα αποτελέσματα της μελέτης, η εφαρμογή αυτών των αλγορίθμων κρίνεται ως αποτελεσματική. Αναφορικά με την μείωση των διαστάσεων των χρονοσειρών, τα αποτελέσματα μας παρέχουν θετικές ενδείξεις σχετικά με την αποτελεσματικότητα και την αποδοτικότητα των αλγόριθμων μείωσης του όγκου των δεδομένων, γεγονός το οποίο μπορεί να αποτελέσει το αντικείμενο μελλοντικής έρευνας.

## ΒΙΒΛΙΟΓΡΑΦΙΑ

Aha, D. W.; Kibler, D. F. & Albert, M. K. (1991), 'Instance-Based Learning Algorithms', Machine Learning 6, 37-66

Agrawal R,. Ghosh S, Imielinski T., Lyer B, Swami A. (1992), An Interval Classifier for Database Mining Applications. In proc of 18th Conf. On Very Large Data Bases (VLDB 1992), pages 560-573

Angiulli, F. (2007), 'Fast Nearest Neighbor Condensation for Large Data Sets Classification', IEEE Trans. on Knowl. and Data Eng. 19(11), 1450-1464

Brighton, H. & Mellish, C. (2002), 'Advances in Instance Selection for Instance-Based Learning Algorithms', Data Min. Knowl. Discov. 6(2), 153-172

Chen, C. H. & Jozwik, A. (1996), 'A sample set condensation algorithm for the class sensitive artificial neural network', Pattern Recogn. Lett. 17, 819-823

Dasarathy B. V. (1991), Nearest neighbor (NN) norms : NN pattern classification techniques. IEEE Computer Society Press

Devi, V. S. & Murty, M. N. (2002), 'An incremental prototype set building technique', Pattern Recognition 35(2), 505-513

Dunham Margaret (2003), Data Mining: Introductory and Advanced Topics. Prentice Hall

Garcia, S., Derrac, J., Cano, J., Herrera, F. (2011), Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence 99 (prePrints)

Gates, G. W. (1972), 'The reduced nearest neighbor rule. IEEE Transactions on Information Theory', IEEE Transactions on Information Theory 18(3), 431-433.

Grochowski, M. & Jankowski, N. (2004), Comparison of Instance Selection Algorithms II. Results and Comments'Artificial Intelligence and Soft Computing - ICAISC 2004', Springer Berlin / Heidelberg, , pp. 580-585

Guttman A. (1984), R-trees: A dynamic index for geometric data. In Proc. ACM SIGMOD International Conf. Management of Data, pages 47-57.

Han, J., Kamber, M., Pei, J. (2011), Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science

Hart, P. E. (1968), 'The condensed nearest neighbor rule', IEEE Transactions on Information Theory 14(3), 515-516.

Jankowski, N. & Grochowski, M. (2004), Comparison of Instances Seletion Algorithms I. Algorithms Survey'Artificial Intelligence and Soft Computing - ICAISC 2004', Springer Berlin / Heidelberg, pp. 598-603

Jolliffe I. T. (2002) Principal Component Analysis, 2nd ed. Springer Series in Statistics

Lam, W.; Keung, C.-K. & Ling, C. X. (2002), 'Learning good prototypes for classification using filtering and abstraction of instances', Pattern Recognition 35(7), 1491 - 1506.

Lozano, M. (2007), Data Reduction Techniques in Classification processes (Phd Thesis). Universitat Jaume I

Manolopoulos Y., Nanopoulos, A. Papadopoulos, A. N., Theodoridis, Y (2006), “R-Tress: Theory and Applications”, Springer

Olvera-López, J. A.; Carrasco-Ochoa, J. A.; Martnez-Trinidad, J. F. & Kittler, J. (2010), 'A review of instance selection methods', Artif. Intell. Rev. 34(2), 133-143

Ritter, G.; Woodruff, H.; Lowry, S. & Isenhour, T. (1975), 'An algorithm for a selective nearest neighbor decision rule', IEEE Trans. on Inf. Theory 21(6), 665-669.

Robinson J. T. (1981). The k-d-b-tree: a search structure for large multidimensional dynamic indexes. In Proceedings of the 1981 ACM SIGMOD international conference on Management of data, SIGMOD

Samet, H. (2006), Foundations of multidimensional and metric data structures. The Morgan Kaufmann series in computer graphics. Elsevier,Morgan Kaufmann

Sanchez, J. S. (2004), 'High training set size reduction by space partitioning and prototype abstraction', Pattern Recognition 37(7), 1561-1564

Sanchez, J.S. (2004): High training set size reduction by space partitioning and prototype abstraction. Pattern Recognition 37(7), 1561–1564

Tomek, I. (1976), 'An experiment with the edited nearest-neighbor rule', IEEE Transactions on Systems, Man, and Cybernetics 6, 448-452

Toussaint, G. (2002), Proximity graphs for nearest neighbor decision rules: Recent progress. In: 34th Symposium on the INTERFACE, pp. 17–20 (2002)

Triguero, I., Derrac, J., Garcıa, S., Herrera, F. (2012), A taxonomy and experimental study on prototype generation for nearest neighbor classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C 42(1), pages 86–100

Wilson, D. L. (1972), 'Asymptotic Properties of Nearest Neighbor Rules Using Edited Data', IEEE trans. on systems, man, and cybernetics 2(3), 408-421

Wilson, D.R., Martinez, T.R. (2000), Reduction techniques for instance-based learning algorithms. Machine Learning 38(3), 257–286

Yianilos P. N. (1993), Data structures and algorithms for nearest neighbor search in general metric spaces. In Proc. of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 311-321

Yi Byoung-Kee and Faloutsos Christos (2000), Fast Time Sequence Indexing for Arbitrary Lp Norms, In Proc. of the 26th International Conference on Very Large Data Bases (VLDB), pages 385-394

Zezula, P., Amato, G., Dohnal, V., Batko, M. (2006), Similarity Search - The Metric Space Approach, vol. 32. Springer, Heidelberg

Keogh Eamonn J., Pazzani Michael J (2000), A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases, In. Proc. of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PADKK 2000), pages 122-133

Keogh Eamonn, Chakrabarti Kaushik, Pazzani Michael, Mehrotra Sharad (2001), Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases, Knowledge and Information Systems (KAIS), Springer-Verlag, pages 263-286

Xi Xiaopeng, Keogh Eamonn, Shelton Christian, Wei Li (2006), Fast Time Series Classification Using Numerosity Reduction, in Proc. of the 23nd International Conference on Machine Learning, Pittsburgh Pages 1033-1040

Buza Krisztian, Nanopoulos Alexandros, and Schmidt-Thieme Lars (2011), INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification, in Proc. of PAKDD 2011, Part II, LNAI 6635, Springer-Verlag, pages 149–160